

System Report for CCL24-Eval Task 9: Bridging the Gap between Authentic and Answer-Guided Images for Chinese Vision-Language Understanding Enhancement

Feiyu Wang¹, Wenyu Guo¹, Dong Yu^{1*}, Chen Kang, Pengyuan Liu^{1,2}

1.Faculty of Computer Science, Beijing Language and Culture University, Beijing, 100083

2.National Language Resources Monitoring and Research Center for Print Media, Beijing, 100083

wfy_0502@163.com, xk17guowenyu@126.com, yudong@blcu.edu.cn

kangchen@blcu.edu.cn, liupengyuan@pku.edu.cn

Abstract

The objective of the Chinese Vision-Language Understanding Evaluation (CVLUE) is to comprehensively assess the performance of Chinese vision-language multimodal pre-trained models in multimodal modeling and understanding across four tasks: Image-Text Retrieval, Visual Question Answering, Visual Grounding, and Visual Dialog. To enhance the models' performance across various multimodal tasks, this paper propose a multimodal information understanding enhancement method based on answer-guided images. Firstly, we propose task-specific methods for answer-guided image generation. Secondly, the authentic and answer-guided images are fed into the model for multimodal fine-tuning, respectively. Finally, training objectives are set for different tasks to minimize the gap between the answer-guided images and authentic images, thereby supervising the results produced by the authentic images utilizing answer-guided images. The experimental results demonstrate the effectiveness of the proposed method.

1 Introduction

The Chinese Vision-Language Understanding Evaluation (CVLUE) aims to assess Chinese vision-language multimodal pre-trained models from multiple perspectives, including Image-Text Retrieval (ITR), Visual Question Answering (VQA), Visual Grounding (VG), and Visual Dialog (VD). This comprehensive evaluation is designed to measure the multimodal modeling and understanding capabilities of these models. Exploring the diverse dimensions of Chinese multimodal pre-trained models not only refines modeling strategies and optimization algorithms, but also significantly enhances the models' ability in multimodal information comprehension and interaction. Researchers can also gain a deeper insight into their practical applications within real-world Chinese contexts.

To enhance the capabilities of Chinese vision-language multimodal pre-trained models across various multimodal tasks, we propose a method to strengthen Chinese multimodal comprehension by bridging the gap between authentic and answer-guided images. Firstly, we propose a method to generate answer-guided images for each task. Specifically, for the ITR task, we generate images according to image captions utilize Chinese text-to-image generation models. The synthetic images effectively highlight key textual information through attributes such as color, quantity, and orientation. For the VQA and VD tasks, we adopt an image generation approach based on questions and answers, integrating answer information into the image generation process. For the VG task, images with grounding information are used as answer-guided images. Secondly, both the answer-guided images and authentic images are fed into the model, respectively. Thirdly, we bridge the gap between authentic and answer-guided images during the training process, thereby enabling the answer-guided images to supervise and elevate the results yielded by authentic images. Experimental results across various tasks demonstrate the effectiveness of our proposed method in enhancing Chinese multimodal understanding.

*Corresponding author: Dong Yu.

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

2 Background

2.1 X^2 -VLM

X^2 -VLM (Zeng et al., 2022a) represents a modular architecture multimodal vision-language model, which is trained through a unified framework to learn multi-grained visual-language alignments. This model enhances its comprehension of weak-correlated image-text pairs by associating visual concepts such as objects, regions, and images with text descriptions, facilitating the execution of various image-text tasks without the need for additional image annotations. Moreover, X^2 -VLM exhibits commendable cross-lingual and cross-domain adaptability. By replacing text encoders tailored for specific languages or domains, it effectively adapts to image-text tasks across different languages and domains without the necessity for related pre-training, demonstrating its potential and flexibility in multimodal applications. Thus, we have selected X^2 -VLM as the pre-trained model for this Chinese image-text multimodal evaluation.

2.2 VisCPM

VisCPM (Hu et al., 2023) is a family of open-source large multimodal models, which support multimodal conversational capabilities (VisCPM-Chat model) and text-to-image generation capabilities (VisCPM-Paint model) in both Chinese and English, achieving the state-of-the-art performance among Chinese open-source multimodal models. VisCPM is trained based on the large language model CPM-Bee with 10B parameters, fusing visual encoder (Muffin) and visual decoder (Diffusion-UNet) to support visual inputs and outputs. In this evaluation, we utilize the VisCPM-Paint model to generate images according to Chinese prompts related to ITR, VQA and VD tasks.

3 Participating System

This section provides a detailed description of the methods and strategies employed in our evaluation. Our method can be divided into two stages: the generation of answer-guided images and the multimodal fine-tuning for each task. To obtain answer-guided images, we design different prompts for text-to-image generation according to different tasks. We also set specific training objectives to accommodate the characteristics of each task during the multimodal fine-tuning stage.

3.1 Answer-guided Image Generation

In the ITR, VQA and VD tasks, we employ the VisCPM-Paint model for image generation and adopt distinct text-to-image generation prompt strategies for different tasks.

As shown in Table 1, for the ITR task, we concatenate the five different captions associated with each image in the training set to form the prompt for text-to-image generation. For the VQA task, as indicated in Table 2, we combine the questions and answers corresponding to each image in the training set using GPT-3.5 (Ouyang et al., 2022). The output sentences of GPT-3.5 are used as prompts for image generation. For the VD task, considering that the majority of images in the training set for the VD task overlap with those in the ITR task, we choose to use the images generated for the ITR task as the answer-guided images for the VD task, while the rest of the images that are not included in the ITR training set are generated using prompts crafted manually from dialogues.

For the VG task, given the particularity of the task, we don't employ the method of text-to-image generation to obtain answer-guided images. As shown in Figure 1, we annotate the bounding boxes of each phrase directly on the authentic images based on the provided bounding box data in the training set. The images are annotated on the authentic images then utilized as the answer-guided images for this task.

3.2 Multimodal Fine-tuning

For different tasks, we design specific training objectives for multimodal fine-tuning. Given the discrepancies in distribution, color, and orientation information between answer-guided images and authentic images, and considering that answer-guided images contain the visual information required by multimodal tasks from text-image pre-trained model, we propose a method to bridge the gap between

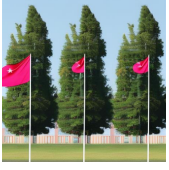
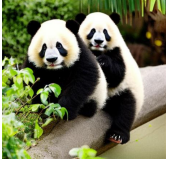

Task	Train Set	Text-to-image Generation Prompt	Generated images
ITR	"caption": ["旗杆上有三面红旗","天空下有三根旗杆, 每根旗杆上都挂着一面红旗", "三面红旗挂在旗杆上, 旗杆下面还有一些树"]	旗杆上有三面红旗, 天空下有三根旗杆, 每根旗杆上都挂着一面红旗, 三面红旗挂在旗杆上, 旗杆下面还有一些树。	
VQA	{ "question": "有几只大熊猫?", "answer": "2"}, { "question": "大熊猫在户外吗?", "answer": "是"}, { "question": "周围有植物吗?", "answer": "有" }	图中共有两只大熊猫, 大熊猫在户外, 周围有植物。	
VD	"dialogues": [{"question": "这个人坐在海边做什么?", "answer": "这个人坐在海边在欣赏风景、放松。"}, {"question": "这个人的表情是怎样的?", "answer": "看不到表情但是她很放松。"}, {"question": "这个人坐在海边身边有什么?", "answer": "身边有海浪, 桌凳, 酒水等其他景观。"}], {"question": "这个人是个男人还是女人?", "answer": "是一个很漂亮的女人。"}]	一个很漂亮的女人坐在海边欣赏风景, 身边有海浪, 桌凳, 酒水等其他景观。虽然看不到表情但是她很放松。	

Table 1: Prompt examples for text-to-image generation.

Prompt	
	请你把下面的问答转化为陈述句: 有几只大熊猫? 3; 有两只熊猫是抱在一起的吗? 是; 大熊猫在人群的哪边? 前边。 图中共有三只大熊猫, 其中有两只是抱在一起的, 而这些大熊猫位于人群的前边。 请你把下面的问答转化为陈述句: 有几只大熊猫? 2; 大熊猫在户外吗? 是; 周围有植物吗? 有。
GPT-3.5	图中共有两只大熊猫, 大熊猫在户外, 周围有植物。

Table 2: Declarative sentence generation for the VQA task.

authentic images and answer-guided images during the training process through targeted training objectives. This approach enables the model to learn from answer-guided images during the training process and generate more qualified results during the inference process without answer-guided images.

For the input text T , authentic image X , and answer-guided image Y , we employ the text encoder to encode the text T into text embedding \mathbf{t} , and we employ a shared-parameter visual encoder to encode the authentic image X and the answer-guided image Y into visual embedding \mathbf{x} and \mathbf{y} respectively. We then define the similarity between the authentic image and the text, as well as the similarity between the answer-guided image and the text, as follows:

$$s(X, T) = g_x(\mathbf{x}_{\text{cls}})^\top g_t(\mathbf{t}_{\text{cls}}), \quad (1)$$

$$s(Y, T) = g_y(\mathbf{y}_{\text{cls}})^\top g_t(\mathbf{t}_{\text{cls}}), \quad (2)$$

where \mathbf{t}_{cls} is the output [CLS] embedding of the text encoder, and \mathbf{x}_{cls} and \mathbf{y}_{cls} are the output [CLS] embedding of the visual encoder. g_t , g_x and g_y are transformations that map the [CLS] embeddings to normalized lower-dimensional representations. Based on it, when the batch size is N , we calculate the in-batch authentic image-to-text and text-to-image similarity as:

$$p^{\text{x2t}}(X) = \frac{\exp(s(X, T)/\tau)}{\sum_{i=1}^N \exp(s(X, T^i)/\tau)}, \quad (3)$$

$$p^{\text{t2x}}(T) = \frac{\exp(s(X, T)/\tau)}{\sum_{i=1}^N \exp(s(X^i, T)/\tau)}, \quad (4)$$

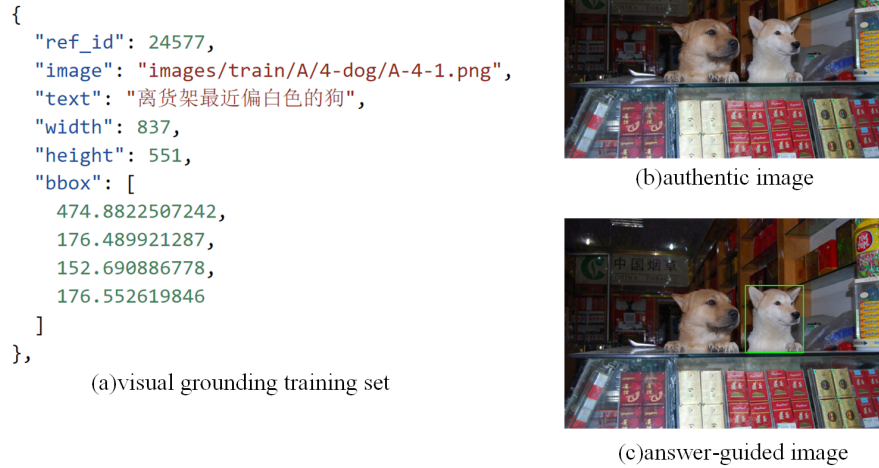


Figure 1: Answer-guided image obtaining for the VG task.

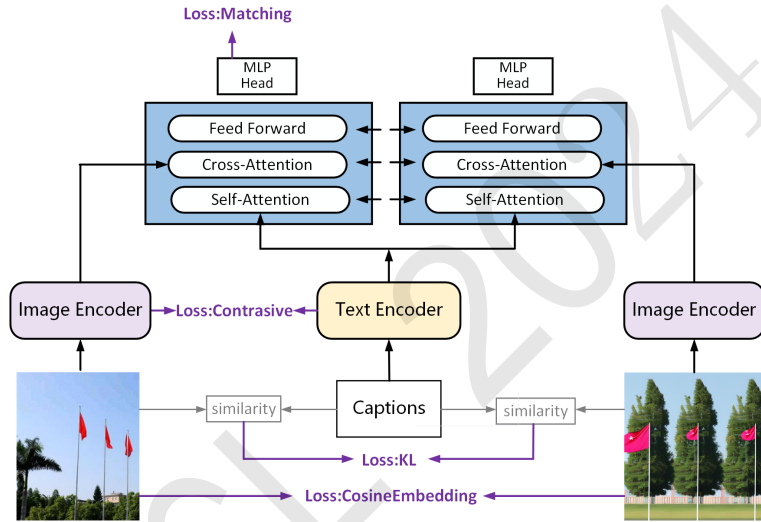


Figure 2: The model architecture of the ITR task.

Similarly, the answer-guided image-to-text similarity is defined as follows:

$$p^{y2t}(Y) = \frac{\exp(s(Y, T)/\tau)}{\sum_{i=1}^N \exp(s(Y, T^i)/\tau)}, \quad (5)$$

where τ is a learnable temperature parameter.

For the ITR task, we use the model architecture in Figure 2. Following previous works (Zeng et al., 2022b; Zeng et al., 2023), in order to align authentic images with texts, we employ contrastive loss as the training objective during fine-tuning. Let $u^{x2t}(X)$ and $u^{t2x}(T)$ denote the ground-truth one-hot similarity, and the contrastive loss is defined as the cross-entropy H between \mathbf{p} and \mathbf{u} :

$$\mathcal{L}_{cl} = \frac{1}{2} \mathbb{E}_{X, T \sim D} [H(\mathbf{u}^{x2t}(X), \mathbf{p}^{x2t}(X)) + H(\mathbf{u}^{t2x}(T), \mathbf{p}^{t2x}(T))], \quad (6)$$

We also utilize the matching loss to ascertain the alignment between an image and its corresponding text:

$$\mathcal{L}_{match} = \mathbb{E}_{X, T \sim D} [H(\mathbf{u}_{match}, \mathbf{p}_{match}(X, T))], \quad (7)$$

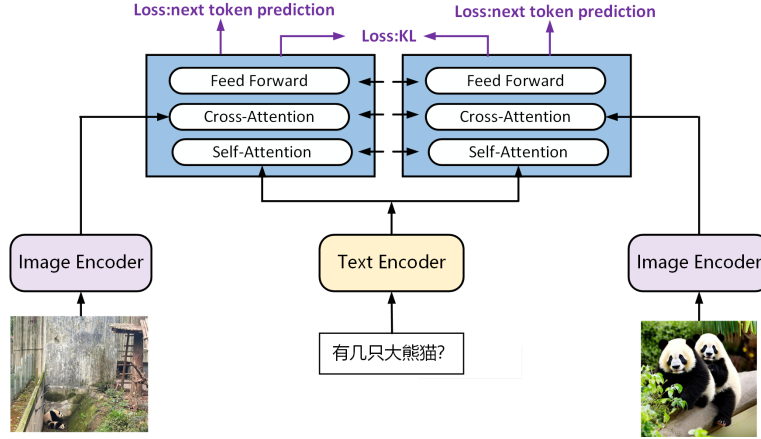


Figure 3: The model architecture of the VQA and VD task.

where $\mathbf{p}_{\text{match}}$ denotes the matching probability of image-text pair predicted by the model, $\mathbf{u}_{\text{match}}$ is a 2-dimensional one-hot vector representing the ground-truth label.

Innovatively, we employ cosine embedding loss and Kullback-Leibler (KL) divergence loss to eliminate the discrepancies between authentic and answer-guided images. Firstly, we use the cosine embedding loss to eliminate the visual representation gap. The loss is calculated as follows:

$$\mathcal{L}_{\text{cos}} = \text{CosineEmbeddingLoss} [g_y(\mathbf{y}_{\text{cls}}) \| g_x(\mathbf{x}_{\text{cls}})], \quad (8)$$

Secondly, because similarity matrices play a key role in calculating the contrastive loss and matching loss, inspired by previous works (Fang and Feng, 2023; Zhou and Long, 2023; Guo et al., 2023; Zhang et al., 2023; Fang et al., 2022), we introduce the Kullback-Leibler (KL) divergence loss to bridge the gap between the authentic and answer-guided image-to-text similarity matrices:

$$\mathcal{L}_1 = \text{KL} [p^{y2t}(Y) \| p^{x2t}(X)], \quad (9)$$

Finally, we employ the contrastive loss, the matching loss, the cosine embedding loss and the KL divergence loss as combined objectives for optimization during the multimodal fine-tuning stage:

$$\mathcal{L}_{\text{itr}} = \mathcal{L}_{\text{cl}} + \mathcal{L}_{\text{match}} + \lambda \mathcal{L}_{\text{cos}} + \gamma \mathcal{L}_1, \quad (10)$$

where λ and γ are hyperparameters that control the contribution of the cosine embedding loss and the KL divergence loss.

For the VQA and VD tasks, as proposed in Figure 3, we utilize the next token prediction loss to train our model. Specifically, we denote the correct answer sentence as $u = (u_1, \dots, u_M)$. The loss functions of the authentic and answer-guided images are calculated respectively:

$$\mathcal{L}_x = - \sum_{j=1}^M \log p(u_j | u_{<j}, t, x), \quad (11)$$

$$\mathcal{L}_y = - \sum_{j=1}^M \log p(u_j | u_{<j}, t, y), \quad (12)$$

Innovatively, because the prediction probabilities have a key impact on calculating the next token prediction loss, we utilize KL divergence loss to enhance the prediction consistency produced by both types of images at the decoder side:

$$\mathcal{L}_2 = \sum_{j=1}^M \text{KL} [\mathbf{p}(u_j | u_{<j}, t, y) \| \mathbf{p}(u_j | u_{<j}, t, x)], \quad (13)$$

Finally, the training objective of the VQA and VD task can be defined as:

$$\mathcal{L} = \mathcal{L}_x + \mathcal{L}_y + \mathcal{L}_2, \quad (14)$$

For the VG task, the fine-tuning model architecture is similar to the VQA or VD task. The bounding box of the given entity i is defined as $\mathbf{b}^i = (lx, ly, w, h)$. The authentic image X and answer-guided image Y are sent into the model to predict the bounding box of the entity, respectively. The predicted bounding boxes are as follows:

$$\hat{\mathbf{b}}_x^i(X, T^i) = \text{Sigmoid}(\text{MLP}(\mathbf{c}_{\text{cls}}^i)), \quad (15)$$

$$\hat{\mathbf{b}}_y^i(Y, T^i) = \text{Sigmoid}(\text{MLP}(\mathbf{c}_{\text{cls}}^i)), \quad (16)$$

where Sigmoid is for normalization, MLP denotes multi-layer perceptron, $\mathbf{c}_{\text{cls}}^i$ is the [CLS] embedding of the fusion module given the features of X (the authentic image) and T (the description of the entity), and $\mathbf{c}_{\text{cls}}^i$ is obtained the same way from Y (answer-guided image) and T .

Following the previous work (Zeng et al., 2022b), we employ the same loss function in the pre-training stage to minimize the discrepancy between the predicted bounding box from the authentic image and the target bounding box:

$$\mathcal{L}_{\text{bbox}} = \mathbb{E}_{(X, T^i) \sim D} \left[L_{\text{iou}}(\mathbf{b}^i, \hat{\mathbf{b}}^i) + \|\mathbf{b}^i - \hat{\mathbf{b}}^i\|_1 \right], \quad (17)$$

We utilize the KL divergence loss to minimize the distance between the predicted bounding boxes generated from authentic and answer-guided images:

$$\mathcal{L}_3 = \text{KL} \left[\hat{\mathbf{b}}_y^i(Y, T^i) \parallel \hat{\mathbf{b}}_x^i(X, T^i) \right], \quad (18)$$

Finally, the training objective of the VG task is as follows:

$$\mathcal{L}_{\text{vg}} = \mathcal{L}_{\text{bbox}} + \mathcal{L}_3. \quad (19)$$

4 Experiment

Dataset We conduct experiments on the dataset provided by the organizer⁰ for both fine-tuning and testing stages. The dataset includes 15 major categories and 92 subcategories of images. The collection of images is carried out manually according to the categories, and there is a strict requirement that the content of the images must be representative of the Chinese cultural environment or commonly seen in daily life.

Pre-trained Models We utilize the CCLM- X^2 VLM-base¹ to initialize our model. We employ BEiT-2 (Peng et al., 2022) as our image encoder and XLM-RoBERTa-base (Conneau et al., 2020) as our text encoder. Additionally, for tasks requiring image-text generation, we employ the VisCPM-Paint model² based on the specified image-text generation prompts.

Systems Settings We set the hyperparameters λ and γ introduced in the ITR task to 0.5, while the rest of parameters follow the parameters set in the baseline model provided by the organizer. In terms of computing resources, we fine-tune the model on 2 V100 for the ITR task and on 4 A100 for tasks involving VQA, VD, and VG during the fine-tuning phrase. During the testing phrase, all tasks are tested on 4 A100 to obtain results.

We utilize the same system setting in the baseline system and our model. The answer-guided images and consistency training objects are removed in the baseline system. The experimental results on the validation datasets are shown in Table 3. By adopting the answer-guided images and our proposed

⁰<https://github.com/WangYuxuan93/CVLUE/tree/main>

¹https://lf-robot-opensource.bytetos.com/obj/lab-robot-public/x2vlm_ckpts_2release/cclm_x2vlm_base.th

²https://huggingface.co/openbmb/VisCPM-Paint/blob/main/pytorch_model.bin

Task	TR			IR			VQA	VD			VG
Metrics	R@1	R@5	R@10	R@1	R@5	R@10	ACC	R@1	R@5	R@10	IoU
Baseline	56.5	83.7	89.8	39.7	67.7	78.4	52.4	28.6	41.2	47.2	48.6
OURS	57.0	83.9	90.4	39.8	68.0	78.8	53.6	28.6	41.4	47.8	49.3

Table 3: The experimental results on the validation dataset.

Task	TR			IR		
Metrics	R@1	R@5	R@10	R@1	R@5	R@10
OURS	57.0	83.9	90.4	39.8	68.0	78.8
-remove KL	55.7	82.1	89.9	38.9	67.0	77.5
-remove COSINE	55.9	82.6	89.6	39.5	67.2	78.2
-replace COSINE with L2	56.6	83.0	90.0	40.0	67.6	78.2

Table 4: Ablation study on different training objectives in the ITR task.

Task	TR			IR			VQA	VD			VG
Metrics	R@1	R@5	R@10	R@1	R@5	R@10	ACC	R@1	R@5	R@10	IoU
OURS	57.0	83.9	90.4	39.8	68.0	78.8	53.6	28.6	41.4	47.8	49.3
-NOISE	52.7	76.9	84.4	39.7	67.2	78.3	52.2	27.4	40.8	47.0	48.9
-RANDOM	53.2	77.9	85.0	39.3	67.4	78.0	52.5	27.4	40.2	46.1	48.4

Table 5: Ablation study on different answer-guided images in all tasks.

consistency training objectives during the fine-tuning phase, the performance across various tasks shows a certain degree of improvement. The improvements show the effectiveness of the answer-guided images. The proposed training objects in bridging the gap between authentic and answer-guided images are able to boost the comprehension of Chinese multimodal contexts.

Ablation Study To further prove the effectiveness of our proposed methods, we conduct the following set of ablation experiments: 1) Ablation study on training objectives in the ITR task; 2) Ablation study on answer-guided images.

As shown in Table 4, we conduct studies on the training objectives in the ITR task to assess the impact of different loss functions on model performance, which involves the removal of the KL divergence loss or the cosine embedding loss, and the substitution of the cosine embedding loss with L2 loss. The drop of results indicates the effectiveness of our proposed training objectives, it also proves the effectiveness of our proposed training method in improving prediction consistency and mitigating the representation disparity.

As shown in Table 5, we conduct studies to assess the effectiveness of answer-guided images generated using our method in all tasks. We compare our model’s performance with two regularization methods: NOISE and RANDOM. NOISE means using noise vectors as the answer-guided image representations. RANDOM means shuffling the correspondences between answer-guided images and textual queries in the training set. We observe a obvious decline using two regularization methods, which shows that the semantic information in the answer-guided images play a key role in our proposed method.

5 Related Work

Multimodal Pre-training Multimodal pre-training research includes the acquisition and clean of large-scale multi-modal data and the design of network architectures and pre-training objectives and so on. We focus on the design of pre-training objectives. CLIP (Radford et al., 2021) is trained on contrastive learning loss, which is widely used in dual-modality. Unicoder-VL (Li et al., 2020) utilizes the visual-linguistic matching loss to extract the positive and negative image-sentence pairs and predict

whether the given sample pairs are aligned or not. Unicoder-VL also uses the masked object classification loss to predict the object category of the masked image regions. UNITER (Chen et al., 2020) uses the word-region alignment loss which targets at explicitly achieves the fine-grained alignment between the multimodal inputs. E2E-VLP (Xu et al., 2021) use the image-text generation model to generate text based on a given image. Ling (Ling et al., 2022) uses the multimodal sentiment prediction loss to enhance the pre-trained models by capturing the subjective information from vision-language inputs. Image-conditioned denoising autoencoding is adopted in XGPT (Xia et al., 2020) to align the underlying image-text using an attention matrix. LXMERT (Tan and Bansal, 2019) uses the masked object regression loss to regress the masked feature or image regions. In our method, we use the contrastive loss, the image-text matching loss, the next token prediction loss and other task-specific prediction losses to train our model.

Chinese Text-to-image Generation The mainstream Chinese diffusion image generation models are derived from further training based on stable-diffusion (Rombach et al., 2022). Some researchers replace the CLIP text encoder with a bilingual encoder or Chinese encoder Taiyi-CLIP (Wang et al., 2022), Chinese-CLIP (Yang et al., 2022), and Alt-CLIP (Chen et al., 2023b), followed by pre-training for text-image matching on a Chinese text-image dataset. Some researchers train on a Chinese text-image dataset for text-to-image generation and obtain the Chinese version of the diffusion image generation model Taiyi-diffusion (Wang et al., 2022) and Alt-diffusion (Ye et al., 2024). ERNIE-ViLG 2.0 (Feng et al., 2023) embarked on training a Chinese diffusion model from scratch using Chinese image-text pairs. In the era of LLMs, PaLI (Chen et al., 2023a) develops a 17B multilingual language-image model based on 10B image-text pairs spanning 100 languages. MultiFusion (Bellagente et al., 2023) discovers that the multilingual language model can help cross-lingual transfer in text-to-image generation. We use the VisCPM (Hu et al., 2023) model, demonstrating that the zero-shot transfer performance of multilingual multimodal models can surpass that of models trained on Chinese multimodal data.

6 Conclusion

In the Chinese Vision-Language Understanding Evaluation task, we propose a system that enhances Chinese text-image multimodal understanding by bridging the gap between authentic images and answer-guided images. Firstly, we propose an image generation module, utilizing the text-to-image generation model or answer information from the train set to generate answer-guided images for different tasks. Secondly, we send the authentic and answer-guided images into the model respectively during the fine-tuning stage. Thirdly, we design specific training objectives for different tasks to encourage the representation or prediction consistency between the two images. Our proposed method improves the performance of our model over the baseline model across all the tasks on the validation set.

Acknowledgements

This work is funded by the Humanity and Social Science Youth foundation of Ministry of Education (23YJAZH184) and the Fundamental Research Funds for the Central Universities in BLCU (No.21PT04).

References

- Marco Bellagente, Manuel Brack, Hannah Teufel, Felix Friedrich, Björn Deiseroth, Constantin Eichenberg, Andrew Dai, Robert Baldock, Souradeep Nanda, Koen Oostermeijer, Andrés Felipe Cruz-Salinas, Patrick Schramowski, Kristian Kersting, and Samuel Weinbach. 2023. Multifusion: Fusing pre-trained models for multi-lingual, multi-modal image generation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: universal image-text representation learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

- Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, and Weicheng Kuo. 2023a. Pali: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Qinghong Yang, and Ledell Wu. 2023b. Altclip: Altering the language encoder in CLIP for extended language capabilities. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8666–8682. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.
- Qingkai Fang and Yang Feng. 2023. Understanding and bridging the modality gap for speech translation. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15864–15881. Association for Computational Linguistics.
- Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. Stemm: Self-learning with speech-text manifold mixup for speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.
- Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiayang Liu, Weichong Yin, Shikun Feng, Yu Sun, Li Chen, Hao Tian, Hua Wu, and Haifeng Wang. 2023. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 10135–10145. IEEE.
- Wenyu Guo, Qingkai Fang, Dong Yu, and Yang Feng. 2023. Bridging the gap between synthetic and authentic images for multimodal machine translation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2863–2874. Association for Computational Linguistics.
- Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038*.
- Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11336–11344. AAAI Press.
- Yan Ling, Jianfei Yu, and Rui Xia. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2149–2159. Association for Computational Linguistics.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. 2022. Beit v2: Masked image modeling with vector-quantized visual tokenizers.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE.
- Hao Tan and Mohit Bansal. 2019. LXMERT: learning cross-modality encoder representations from transformers. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5099–5110. Association for Computational Linguistics.
- Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, Chongpei Chen, Ruyi Gan, and Jiaying Zhang. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.
- Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, and Ming Zhou. 2020. Xgpt: Cross-modal generative pre-training for image captioning.
- Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. 2021. E2E-VLP: end-to-end vision-language pre-training enhanced by visual learning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 503–513. Association for Computational Linguistics.
- An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese CLIP: contrastive vision-language pretraining in chinese. *CoRR*, abs/2211.01335.
- Fulong Ye, Guang Liu, Xinya Wu, and Ledell Wu. 2024. Altdiffusion: A multilingual text-to-image diffusion model. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pages 6648–6656. AAAI Press.
- Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. 2022a. X²-vlm: All-in-one pre-trained model for vision-language tasks. *CoRR*, abs/2211.12402.
- Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. 2022b. X²-vlm: All-in-one pre-trained model for vision-language tasks. *CoRR*, abs/2211.12402.
- Yan Zeng, Wangchunshu Zhou, Ao Luo, Ziming Cheng, and Xinsong Zhang. 2023. Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5731–5746. Association for Computational Linguistics.
- Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shang-tong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models.
- Yucheng Zhou and Guodong Long. 2023. Improving cross-modal alignment for text-guided image inpainting. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3445–3456, Dubrovnik, Croatia, May. Association for Computational Linguistics.