UniGen: Unified Generative Pre-training for Multilingual Multimodal Representation

Zheyuan Tian
State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China, School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
tianzheyuan2021@ia.ac.cn

Guan Luo*
State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China, School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
gluo@nlpr.ia.ac.cn

Bo Wang
State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China, School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
wangbo@ia.ac.cn

Bing Li
State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China, School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
bli@nlpr.ia.ac.cn

Weiming Hu
State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China, School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China , School of Information Science and Technology, ShanghaiTech University, Shanghai, China
wmhu@nlpr.ia.ac.cn

ABSTRACT

Multilingual multimodal pre-training has garnered significant attention, but it faces challenges due to the substantial need for diverse multilingual text-image data, especially for minor languages. This article introduces UniGen, a unified strategy for efficient multilingual multimodal pre-training inspired by internet data distribution observations. Leveraging the richer availability and higher quality of multilingual text-English text and English text-image data, UniGen aligns the latent space of multilingual text with visual information to a unified semantic space. This alignment, with English as a reference, proves effective in enhancing cross-modal understanding. UniGen reduces reliance on multilingual text-image data, surpassing comparable models in multilingual multimodal benchmark IGLUE by a notable 7%. Notably, UniGen is the first multilingual multimodal model to unify all pre-training tasks within a generative pre-training framework.

*Corresponding author

ICIAI 2024, March 16–18, 2024, Tokyo, Japan

ACM ISBN 979-8-4007-0930-2/24/03
https://doi.org/10.1145/3655497.3655509

CCS CONCEPTS

• **Computing methodologies** → Machine learning; Machine learning approaches; Neural networks; Artificial intelligence; Computer vision; Computer vision representations; Image representations; Artificial intelligence; Natural language processing; Natural language generation.

KEYWORDS

Multimodal pre-training, Multilingual model, autoregressive model, generative model

ACM Reference Format:
Zheyuan Tian, Guan Luo, Bo Wang, Bing Li, and Weiming Hu. 2024. UniGen: Unified Generative Pre-training for Multilingual Multimodal Representation. In *2024 the 8th International Conference on Innovation in Artificial Intelligence (ICIAI 2024), March 16–18, 2024, Tokyo, Japan*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3655497.3655509

1 INTRODUCTION

Current research focuses on multilingual and multimodal models, which empower models to comprehend multiple languages and bridge the gap between vision and language. These models are crucial for multilingual communities, enabling tasks like creating image descriptions in various languages. They require sophisticated alignment across different language embeddings and between language and visual data encodings. Such models can process both visual and textual content in multilingual and multimodal contexts.

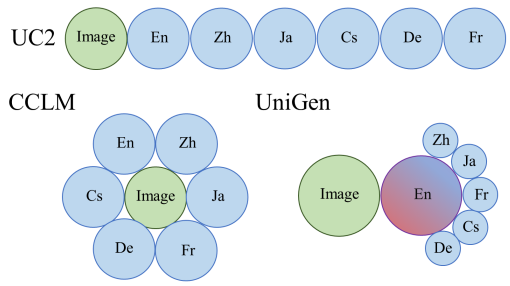


Figure 1: A simple illustration of data reliance. While UC2 requires multilingual all-aligned text-image data, CCLM requires multilingual text-image data, UniGen can rely on English text-image data to achieve multilingual multimodal alignment.

Earlier multilingual multimodal pre-training research has explored different methods. M_3P [1] and UC^2 [2] pioneered using aligned image-text data in multiple languages for pre-training, with specific training tasks and objectives. However, their reliance on closely matched data limits language diversity. To address this, CCLM [3] approach reduces the need for aligned data by treating cross-modality and cross-language as separate views. It leverages images with descriptions in one language, combined with multilingual text, to improve text alignment across languages. However, as illustrated in Figure 1, despite the relatively flexible requirements compared to previous works, this approach still relies on data composed of multilingual text and images, which restricts the potential for data augmentation.

Moreover, the majority of existing pre-training strategies [2–4] often focus on masked language modeling which forces the model to complete sentences by masking certain parts of the text, prompting it to extract information from visual features or forcibly establishing associations between visual targets and text [1, 2, 5] using object detection frame–works like Fast R-CNN [6]. Neither of these approaches directly use visual features

effectively; instead, they serve as auxiliary alignments in the training process for text tasks. The singularity of training objectives prevents the model from exploring deep information in both text and images. Masked data modeling is not specifically designed for cross-lingual and cross-modal tasks. Some efforts attempting to leverage multimodal information also introduce multiple training objectives, increasing training complexity.

In this work, we propose UniGen based on generative pre-training to address the aforementioned issues. On the one hand, at the data level, we simplify data requirements by using English as a pivot between multilingual text and images, reducing dependency on paired data. On the other hand, our unified generative approach streamlines pre-training objectives, treating multilingual-to-English text as translation and English text-to-image as captioning, both handled by a generative decoding model. As noted in BEIT-3 [7], introducing multiple pre-training objectives can be scaling up unfriendly and inefficient for training the model. A unified pre-training paradigm provides certain advantages for model training in this context. Additionally, we integrate traditional tasks like contrastive learning, image-text matching, and question-answering into our generative framework. Drawing inspiration from models like ChatCaptioner [8], we exploit large language models' in-context learning to generate question-answering data, enhancing the model's detail perception without extra datasets. This strategy aims to refine the model's visual and textual understanding.

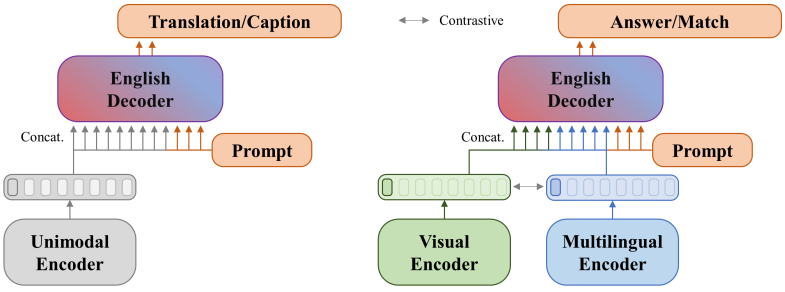
We tested the UniGen model using the IGLUE [9] benchmark to assess its effectiveness. Experiments show that UniGen matches or surpasses current leading methods in comprehension, question-answering, and reasoning, even with limited or no aligned multilingual text-image data. Notably, UniGen excels in understanding minor languages, promoting fairness in model performance. Unlike some previous models that require complex adjustments for different tasks, UniGen employs a consistent decoding approach built on a unified pre-training strategy, simplifying fine-tuning and improving both efficiency and ease of use.

Our contributions can be outlined in three folds: 1) the proposal of a novel training method, UniGen, which employs a unified generative pre-training paradigm for multilingual multimodal pre-training, reducing the data requirements of traditional methods; 2) the utilization of Image Captioning tasks and large-model-based image-text question-answering tasks to more directly leverage visual information to improve model performance; 3) the adoption of a paradigm that unifies pre-training and downstream task transfer training, making the transfer process more straightforward without the need for introducing additional network components. We believe this paradigm offers a new perspective for existing research in multilingual multimodal pre-training, especially in the understanding of multimodal content for minor languages.

2 RELATED WORKS

Pre-trained Language Models In recent years, many pre-trained language models have achieved significant performance leaps by leveraging vast amounts of data, relying on the self-attention mechanism of Transformer [10] and the use of self-supervised tasks. On one hand, the representative of autoencoding pre-training paradigms, BERT [11], achieved excellent text representations through masked language modeling and next-sentence prediction. Subsequently, RoBERTa [12] enhanced robust optimization by employing dynamic masking. These studies have led to the extension of multilingual applications. mBERT [11], trained on a larger multilingual vocabulary and masked language modeling with multilingual text, achieved encoding for multilingual text. XLM-RoBERTa [13] introduced translation language modeling, enhancing semantic alignment across different languages. On the other hand, representative of autoregressive pre-training, the GPT series [14–16] demonstrated that generative pre-training can also achieve language understanding. Furthermore, it explored how generative pre-training models can perform various tasks on top of language understanding and acquire the capability of in-context learning.

Multimodal Pre-training: from mono to multilingual In the field of vision-language pre-training, early researches [17–19] often used a single-stream architecture, which allowed multimodal features to interact early in the network, improving alignment.



a) Unimodal Input b) Multimodal Input

Figure 2: Illustration of the UniGen framework. UniGen involves two types of inputs: (a) unimodal or (b) multimodal input.

This led to the development of pre-training tasks like masked image modeling and image-guided masked language modeling. CLIP [20], with its contrastive learning on a vast image-text dataset, showed remarkable zero-shot learning abilities. More recent studies have combined contrastive learning with feature fusion techniques. ALBEF [21] introduced an 'align before fusion' method, while BLIP-2 [22] used a Q-former to extract visual features relevant to the text, both methods improving multimodal data alignment. In multilingual multimodal pre-training, M3P [1] leveraged multilingual multimodal inputs with random replacement and code-switch tasks to enhance multilingual capabilities. UC2 [2] used dual-language captions and visual data, employing offline models to associate image targets with text and co-masking texts to help the model learn from images. CCLM [3] treated image-multilingual text and multilingual-English text as separate views, using masked language modeling for alignment. This approach led to state-of-the-art performance in various multilingual multimodal tasks.

3 METHOD

3.1 Overview

UniGen framework aims to achieve cross-language and cross-modal alignment while reducing the usage of multilingual text-image data. UniGen comprises three components: an image encoder, a multilingual encoder, and an English decoder. The image encoder consists of a pre-trained BLIP-2_{base} [22] model, which includes a frozen CLIP ViT-L/14 as encoder, and a Q-Former with 32 learnable queries. For an image with a resolution of 224×224 as input, the image encoder first partitions and extracts features in a patch size of 16×16, followed by interaction with learnable queries in the Q-Former to obtain the output. For the text encoder, we employ a pre-trained XLM-R, which has been pre-trained on over 100 languages, as the text encoder. It can tokenize input from multilingual text and extract features to generate output. As for the English decoder, a GPT-2_{base} is chosen as the decoder. It takes the output from the image encoder and the multilingual encoder as input embeddings and produces the final output sequence through the decoder.

In the training tasks, we have selected two categories, totaling four tasks, as illustrated in Figure 2. These comprise two unimodal input tasks and two multimodal input tasks. The unimodal input tasks involve a translation task using multilingual text-English text input and an image captioning task using English text-image input. As for the multimodal input, there is a multimodal question-answering task (details will be introduced in section 3.2 and an image-text matching task. The specifics of the training tasks will be elaborated in section 3.3.

3.2 Large Language Model as Questioner

Masked language modeling (MLM) offers distinct benefits over generative pre-training, particularly in its ability to prompt models to fill in randomly masked tokens, thereby honing the model's attention to crucial caption details and enhancing perception of visual nuances. However, generative autoregressive training lacks this attention on detail within feature representations, which can limit fine-grained perceptual skills. To address the issues of missing image detail and small targets in generative pre-training, we introduce image-text question-answering to encourage the model utilize detailed visual information. Inspired by ChatCaptioner [8] framework, which uses large language models for linguistic guidance in multimodal tasks, our approach employs an offline large language model to extract details from captions, directing multilingual multimodal pre-training.

As depicted in Figure 3, considering the model's size, accessibility, and computational constraints, we utilizes an offline large language model, Llama-30B [25], to extract detailed information from image descriptions in Conceptual Captions 3M. Capitalizing on the in-context learning capability of LLM, prompts are employed to provide examples for the model's output. The specific prompt is as follows:

*Extract accurate and **important** information to form a question-answer pair **exactly from the sentence** following the pattern, **avoid using yes/no answers**:*

<sentence> a worker sits on a bench in front of a group of young people. <question> What is the worker in front of? <answer> young people. <sentence>. . .

Three key prompts are provided to the large language model: 1) Extract important information; 2) Avoid illusions; 3) Avoid simple questions. Through these prompts, detailed information can be continuously extracted from input captions, automatically constructing image-question-answer data for subsequent image-text question-answering training.

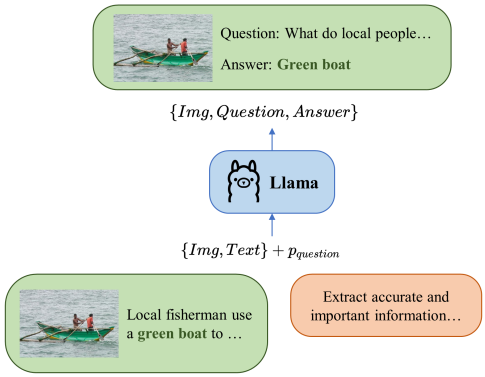


Figure 3: A simple illustration of large language models Questioner.

3.3 UniGen Pre-training

UniGen pre-training process consists of two types: unimodal-input pre-training and multimodal-input pre-training.

Unimodal-input Pre-training The primary objective of unimodal-input pre-training is to decouple multilingual text-image data, achieving alignment between multiple languages and visual latent spaces through a shared English decoder. Although they have different input modalities, both of the final outputs are corresponding English translations or captions, enabling a unified pre-training objective. We additionally introduce two special

tokens, [trans] and [cap], to serve as prompts for task identification in the English decoder.

Taking multilingual-English text pairs as an example, after SentencePiece tokenization and XLM-R feature extraction, the output E_t is then concatenated with [trans] and fed into the decoder, resulting in an output denoted as $D_t^N = [D_0^N, \ldots, D_n^N]$. We employ cross-entropy loss for generative pre-training, translation and caption loss are denoted as $\mathcal{L}_{trans}$ and $\mathcal{L}_{cap}$ respectively.

$$D_{trans}^0 = [E_{t0}, E_{t1}, \ldots, E_{tn}] \oplus T_D\left([trans]\right)$$

$$\mathcal{L}_{trans} = -\frac{1}{N}\sum_{i=0}^{n}\sum_{j=0}^{V} y_{i,j} \log\left(D_{i,j}^N\right)$$

where T_D denotes the decoder tokenizer, $\oplus$ is concatenate operation, V is the vocabulary size and $D_{i,j}^N$ is the predicted probability of token j by the model at position i.

Multimodal-input Pre-training While unimodal input pre-training achieves cross-modal alignment through a shared English decoder, it can be troublesome to achieve better alignment without cross-modal information interaction. Hence, the introduction of multimodal input pre-training aims to enhance cross-modal alignment. This includes two pre-training tasks: image-text question-answering and image-text matching. Correspondingly, two special tokens are introduced: [qus] and [match]. Regardless of which task, the input consists of English text-image pairs (or multilingual text-image pairs). After passing through the multilingual encoder and visual encoder, the embeddings are obtained, denoted as E_v and E_t respectively. After concatenation with the corresponding special token, they are fed to the decoder, producing the output corresponding to the answer or judgment. Cross-entropy loss is used and labels $\mathcal{L}_{qus}$ and $\mathcal{L}_{match}$ are assigned for these tasks.

$$D_{qus}^0 = [E_{v0}, \ldots, E_{vm}] \oplus [E_{t0}, \ldots, E_{tn}] \oplus T_D\left([qus]\right)$$

$$\mathcal{L}_{qus} = -\frac{1}{N}\sum_{i=0}^{n}\sum_{j=0}^{V} y_{i,j} \log\left(D_{i,j}^N\right)$$

Following ALBEF and later works [21, 23], we adopt contrastive learning in the image-text matching task to obtain hard negatives for the matching process, which is vital for cross-modal alignment. Simultaneously, contrastive learning can provide supervision signals for alignment at the model's encoder levels, assisting in cross-modal alignment. It's important to note that, unlike BLIP-2 [22], we don't calculate the similarity for each query and select the maximum similarity as the [CLS] token. Instead, we take the first token from the learnable queries of length m as the [CLS] token, which carries global information of the visual input.

For a given positive sample of an image-text pair v_i, t_i, where the remaining texts t_j and images $v_j (j \neq i)$ serve as negative samples for v_i and t_i respectively, the NCE loss is defined as follows:

$$\mathcal{L}_{CL} = \frac{1}{2}\left(\mathcal{L}_{v2t} + \mathcal{L}_{t2v}\right)$$

$$\mathcal{L}_{v2t} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\text{sim}\left(v_i, t_i\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}\left(v_i, t_j\right)/\tau\right)}$$

where B is the batch size, $\text{sim(a, b)} = \frac{\mathbf{a}^{\mathrm{T}}\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$ is the cosine similarity between two tokens, τ is a learnable hyperparameter that controls the scale of similarity logits.

To sum up, the loss for all tasks consists of four generative losses and one contrastive learning loss:

$\mathcal{L}_{sum} = \mathcal{L}_{trans} + \mathcal{L}_{cap} + \alpha\mathcal{L}_{qus} + \beta\mathcal{L}_{match} + \mathcal{L}_{CL}$. The first two align at the decoder level through single-modal tasks, while the remaining three provide refined alignment objectives for the model through cross-modal tasks.

4 EXPERIMENT

4.1 Datasets

The dataset used in this study consists of a multilingual text-image dataset (primarily in English) and a multilingual-English text dataset. Following the approach in CCLM [3], Conceptual Captions 3M(CC3M) [24] dataset is utilized as the primary multimodal dataset. In UC2, CC3M has been translated into five languages (Czech, French, German, Japanese, and Chinese) to form CC6L. This expansion covers a variety of major languages and has been widely applied in multilingual multimodal pre-training research.

However, the extensive use of data from major languages has to some extent affected the alignment of non-participating minor languages, leading to a decrease in their performance. Therefore, this study attempts to mitigate this issue by reducing the utilization of multilingual multimodal data. Specifically, during pre-training, we randomly selected 10% of CC6L data and used randomly selected language texts and images for input. The remaining portion of the data was exclusively comprised of English data from CC3M for training.

In addition, as described in section 3.2, we introduced an English image-text question-answering dataset through Llama-30B [25]. To ensure fairness, we also utilized CC3M as the source for extracting key information from captions, without introducing additional datasets. This dataset can be referred to as Conceptual Captions Question Answering 3M (CCQA3M).

Finally, for multilingual-English text, we employed the same WikiMatrix [26] subset as CCLM for training. This subset includes 19 million examples of multilingual text pairs in English and 20 languages, covering all languages in the IGLUE benchmark.

4.2 Implementation Details

UniGen contains a visual encoder with CLIP ViT-L/14 and a Q-Former, a XLM-R multilingual text encoder, and a GPT-2_{base} decoder. For efficiency, the ViT part of the visual encoder and the intermediate 6 layers of the decoder are frozen during training. The model has an embedding dimension of 768, and the model consists of 423M trainable parameters in total.

During pre-training, we employed the same 224×224 image input size as in BLIP-2, along with learnable queries of length 32. For XLM-R, a maximum text length of 40 was used. UniGen pre-training uses the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$ and a weight decay of 0.05. The learning rate was set to 1e-4, with a warm-up of 2000 steps and linear decay. Pre-training was conducted with mixed precision on 8 NVIDIA A100 GPUs for 30 epochs, with a batch size of 512, and a training duration of $\sim$ 6 days.

On the fine-tuning phase, we employed a reduced learning rate to fine-tune the decoder model. Throughout this process, the study utilized the small-sample dataset provided by iGLUE for fine-tuning training. For classification tasks, such as XVNLI, we trained the model to output corresponding keywords. For tasks like xGQA, we allowed the model to generate answers from the entire vocabulary. Admittedly, this approach is more challenging compared to the previous research that used a restricted vocabulary for classification. However, we believe that it is a more scalable method and a better reflection of the model's capabilities. During this process, we also encountered instances where the model output "black cat" while the answer was "cat". Such cases were adjudicated manually when compiling the final results. We use an AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$ and a learning rate of 3e-5 without warming up, the fine-tuning process was conducted with mixed precision on 1/8 (for retrieval) NVIDIA A100 GPU for 10 epochs.

4.3 Experimental Results

4.3.1 Downstream tasks. To facilitate a fair comparison with previous works, we validated UniGen on the IGLUE benchmark. The IGLUE benchmark encloses a diverse set of tasks, including classification, question answering, inference, and retrieval, across 20 major and minor languages. Additionally, it provides few-shot data to assess the model's transferability under a computational-efficient scenario.

XVNLI Cross-lingual Visual Natural Language Inference (XVNLI) tasks the model to determine if a given text-hypothesis aligns with, contradicts, or is neutral to an image-premise. The dataset is based on SNLI [27] with the addition of multi-modal [28] and 4 target language counterparts [29].

xGQA Cross-lingual Grounded Question Answering tasks [4] the model to answer questions based on structured questions about a related image. The dataset is manually translated into 7 target languages based on GQA [30] dataset that was originally sampled from Visual Genome.

MaRVL Multicultural Reasoning over Vision and Language dataset [32] requires the model to classify whether the description is true or false about input images. The task relies on hand-written descriptions provided in 5 languages, utilizing NLVR2 [31] for training and MaRVL exclusively for testing.

Retrieval IGLUE comprises two retrieval tasks, namely xFlickr&CO and WIT. For the former, 1000 image-text pairs were selected from the Flickr30K [34] and COCO [35] datasets, respectively, to form the test set. The textual descriptions were manually captioned by annotators following the guidelines of Flickr30K in 6 languages. For WIT, data was gathered from Wikipedia in 108 languages [33], and test sets were constructed in 10 different languages, maintaining a diverse range of language sizes.

4.3.2 IGLUE Results. Table 1 presents UniGen's results on the IGLUE benchmark, detailing its zero-shot and few-shot performance. For retrieval tasks, Recall@1 (R@1) is reported, while accuracy is used for understanding tasks like XVNLI, xGQA, and MaRVL. In the zero-shot scenario, models are fine-tuned using English datasets and tested on various target languages. In the few-shot scenario, fine-tuning occurs on datasets in the target languages. To maintain a fair comparison, we ensured that all models were evaluated using equivalent data volumes and similar model sizes. UniGen shows a significant 7% improvement in zero-shot understanding tasks, including inference and question-answering, despite using only 10% of the multilingual-image data required by previous models. For retrieval tasks, UniGen outperforms CCLM in dual-stream retrieval (precentesd in parentheses), an improvement attributed to a more powerful back-end that enhances performance in rapid retrieval scenarios.

In the few-shot scenario, we employed the same settings as IGLUE, conducting continuous fine-tuning of the model with a limited amount of small samples. For WIT, given the absence of corresponding few-shot tasks in the original benchmark, we adopted a similar setup. It can be observed that, overall, the model benefits from

the learning process with a small number of samples from low-resource languages. The extent of improvement, however, is influenced by both the alignment of the model itself and the quality of the small-sample data.

4.4 Ablation Study

To validate the effectiveness of each component in the model, we conducted ablation studies by removing certain pre-training components to evaluate their contributions to model training. We examined the effects under three conditions: using full CC6L dataset, using only 10% of the data (reported configuration), and not using multilingual text-image data. This setup aims to investigate the contribution of multilingual text-image data to model alignment. Additionally, we explored the effect of different pre-training strategies by reducing question answering data, contrastive learning, and other components on model performance. Apart from the specific

Table 1: Results on IGLUE benchmark

Model	NLI	QA	Reasoning	Retrieval			
	XVNLI	xGQA	MaRVL	xFlickr&CO		WIT	
				IR	TR	IR	TR
Zero-shot setting							
xUNITER	58.48	21.72	54.59	14.04	13.51	8.72	9.81
M_3P	58.25	28.17	56.00	12.91	11.90	8.12	9.98
UC^2	62.05	29.35	57.28	20.31	17.89	7.83	9.09
$CCLM_{3M}$	74.64	42.36	65.91	**67.35**(42.39)	**65.37**(43.04)	**27.46**	**28.66**
Ours	**75.02**	**45.33**	**67.13**	62.59(**43.72**)	60.50(**44.54**)	20.36	20.89
Few-shot setting							
xUNITER	60.55	40.68	57.46	14.30	13.54	-	-
M_3P	59.36	41.04	49.79	13.21	12.26	-	-
UC^2	63.68	42.95	58.32	19.79	17.59	-	-
$CCLM_{3M}$	**75.15**	50.94	70.53	**66.04**	**68.15**	-	-
Ours	74.98	**54.19**	**72.27**	63.70	61.09	-	-
Best Results on Translate-English test							
VisualBERT	74.12	48.72	62.35	41.64	36.44	15.36	15.75
VL-BERT	73.86	49.78	64.16	38.18	31.84	15.11	16.09

Table 2: Ablation study results

Settings	XVNLI	xGQA	MaRVL	xFlickr&CO	
				IR	TR
Ours -w/ 10% CC6L	**75.02**	**45.33**	**67.13**	**62.59**	**60.50**
-w/ 100% CC6L	75.21	46.21	67.48	64.41	63.27
-w/o CC6L	74.80	44.47	64.86	60.58	58.12
-w/o CCQA3M	75.17	43.92	66.91	58.26	58.07
-w/o CCQA3M+CC6L	74.45	43.41	64.49	57.20	56.88
-w/o CCQA3M+CC6L+CL	67.38	40.31	62.10	50.08	47.22

aspects under investigation, we maintained an identical experimental setup to ensure fairness in the comparisons. The results are shown in Table 2.

Firstly, regarding the utilization of multilingual text-image data, we observed that having more such data does contribute to an overall performance improvement. However, this enhancement is primarily confined to the performance improvement of the 6 major languages restricted by CC6L, and the performance of minor languages tends to be adversely affected. without CC6L, UniGen's approach is still capable of achieving alignment effectively through English decoder latent space, yielding performance comparable to that achieved with multilingual text-image data.

Image-text question answering plays a crucial role in improving multilingual question answering and retrieval tasks during pre-training. It sharpens the model's attention to details and enhances its ability to differentiate between similar samples. Eliminating contrastive learning leads to a significant 10% drop in performance, underscoring the importance of hard negative samples in teaching the model to establish stronger connections and alignment, which is vital for the success of downstream comprehension and retrieval tasks.

5 CONCLUSION

In this paper, we propose UniGen, a unified generative pre-training paradigm for multilingual and multimodal pre-training. It decouples the requirement for multilingual text-image data into two distinct data dependencies: multilingual-English text and English text-image. Under the unified UniGen pre-training framework, we achieve pre-training and fine-tuning for downstream tasks. We validate UniGen's pre-training on the IGLUE benchmark, demonstrating stronger multilingual multimodal understanding capabilities with comparable retrieval performance under the premise of using fewer multilingual text-image data. Importantly, our study reduces the data dependency on multilingual text-image by decoupling the data, thereby exhibiting greater potential to achieve improved results on larger datasets. This part remains for future exploration.

ACKNOWLEDGMENTS

This work is supported by the national key R&D program of China (No. 2020AAA0106800), the Natural Science Foundation of China (Grant No. 62036011, 62172413, 62192782, 61721004, U2033210), Beijing Natural Science Foundation (L223003, 4234086), the Project of Beijing Science and technology Committee (Project

No. Z231100005923046) , the Major Projects of Guangdong Education Department for Foundation Research and Applied Research (Grant: 2017KZDXM081, 2018KZDXM066), Guangdong Provincial University Innovation Team Project (Project No.: 2020KCXTD045).

REFERENCES

[1] M. Ni *et al.*, 'M3p: Learning universal representations via multitask multilingual multimodal pre-training', in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 3977–3986.
[2] M. Zhou *et al.*, 'Uc2: Universal cross-lingual cross-modal vision-and-language pre-training', in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4155–4165.
[3] Y. Zeng *et al*, 'Cross-View Language Modeling: Towards Unified Cross-Lingual Cross-Modal Pre-training', in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 5731–5746.
[4] J. Pfeiffer *et al.*, 'xGQA: Cross-lingual visual question answering', arXiv preprint arXiv:2109. 06082, 2021.
[5] P.-Y. Huang, J. Hu, X. Chang, and A. Hauptmann, 'Unsupervised Multimodal Neural Machine Translation with Pseudo Visual Pivoting', in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8226–8237.
[6] R. Girshick, 'Fast r-cnn', in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.
[7] W. Wang *et al.*, 'Image as a Foreign Language: BEIT Pretraining for Vision and Vision-Language Tasks', 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 19175–19186, 2023.
[8] D. Zhu, J. Chen, K. Haydarov, X. Shen, W. Zhang, and M. Elhoseiny, 'Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions', arXiv preprint arXiv:2303. 06594, 2023.
[9] E. Bugliarello *et al.*, 'IGLUE: A benchmark for transfer learning across modalities, tasks, and languages', in International Conference on Machine Learning, 2022, pp. 2370–2392.
[10] A. Vaswani *et al.*, 'Attention is All you Need', in Advances in Neural Information Processing Systems, 2017, vol. 30.
[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', in North American Chapter of the Association for Computational Linguistics, 2019.
[12] Y. Liu *et al.*, 'Roberta: A robustly optimized bert pretraining approach', arXiv preprint arXiv:1907. 11692, 2019.
[13] A. Conneau *et al.*, 'Unsupervised Cross-lingual Representation Learning at Scale', in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.
[14] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, and Others, 'Improving language understanding by generative pre-training', 2018.
[15] A. Radford *et al.*, 'Language models are unsupervised multitask learners', OpenAI blog, vol. 1, no. 8, p. 9, 2019.
[16] T. Brown *et al.*, 'Language Models are Few-Shot Learners', in Advances in Neural Information Processing Systems, 2020, vol. 33, pp. 1877–1901.
[17] G. Li, N. Duan, Y. Fang, D. Jiang, and M. Zhou, 'Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training', in AAAI Conference on Artificial Intelligence, 2019.
[18] Y.-C. Chen *et al.*, 'Uniter: Universal image-text representation learning', in ECCV, 2020.
[19] W. Kim, B. Son, and I. Kim, 'ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision', in International Conference on Machine Learning, 2021.
[20] A. Radford *et al.*, 'Learning Transferable Visual Models From Natural Language Supervision', in International Conference on Machine Learning, 2021.
[21] J. Li, R. R. Selvaraju, A. D. Gotmare, S. R. Joty, C. Xiong, and S. C. H. Hoi, 'Align before Fuse: Vision and Language Representation Learning with Momentum Distillation', in Neural Information Processing Systems, 2021.
[22] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, 'BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models', ArXiv, vol. abs/2301.12597, 2023.
[23] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, 'BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation', in International Conference on Machine Learning, 2022.
[24] P. Sharma, N. Ding, S. Goodman, and R. Soricut, 'Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning', in Proceedings of ACL, 2018.
[25] H. Touvron *et al.*, 'LLaMA: Open and Efficient Foundation Language Models', arXiv [cs.CL]. 2023.
[26] H. Schwenk, V. Chaudhary, S. Sun, H. Gong, and F. Guzmán, 'WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia', in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021, pp. 1351–1361.
[27] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, 'A large annotated corpus for learning natural language inference', in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 632–642.
[28] N. Xie *et al*, 'Visual entailment: A novel task for fine-grained image understanding', ArXiv preprint, vol. abs/1901.06706, 2019.
[29] Ž. Agić and N. Schluter, 'Baselines and Test Data for Cross-Lingual Inference', in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.
[30] D. A. Hudson *et al*, 'GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering', in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, 2019, pp. 6700–6709.
[31] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi, 'A Corpus for Reasoning about Natural Language Grounded in Photographs', in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 6418–6428.
[32] F. Liu, E. Bugliarello, E. M. Ponti, S. Reddy, N. Collier, and D. Elliott, 'Visually Grounded Reasoning across Languages and Cultures', in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 10467–10485.
[33] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. 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, 2021, pp. 2443–2449.
[34] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, 'From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions', Transactions of the Association for Computational Linguistics, vol. 2, pp. 67–78, 2014.
[35] T.-Y. Lin *et al.*, 'Microsoft COCO: Common Objects in Context', in European Conference on Computer Vision, 2014.