RC³: Regularized Contrastive Cross-lingual Cross-modal Pre-training

Chulun Zhou[1]*, Yunlong Liang[2]*, Fandong Meng[1]†, Jinan Xu[2], Jinsong Su[3] and Jie Zhou[1]

[1]Pattern Recognition Center, WeChat AI, Tencent Inc, China
[2]Beijing Key Lab of Traffic Data Analysis and Mining,
Beijing Jiaotong University, Beijing, China
[3]School of Informatics, Xiamen University, Xiamen, China
{chulunzhou,fandongmeng,withtomzhou}@tencent.com
{yunlongliang,jaxu}@bjtu.edu.cn
jssu@xmu.edu.cn

Abstract

Multilingual vision-language (V&L) pre-training has achieved remarkable progress in learning universal representations across different modalities and languages. In spite of recent success, there still remain challenges limiting further improvements of V&L pre-trained models in multilingual settings. Particularly, current V&L pre-training methods rely heavily on strictly-aligned multilingual image-text pairs generated from English-centric datasets through machine translation. However, the cost of collecting and translating such strictly-aligned datasets is usually unbearable. In this paper, we propose **R**egularized **C**ontrastive **C**ross-lingual **C**ross-modal (RC³) pre-training, which further exploits more abundant weakly-aligned multilingual image-text pairs. Specifically, we design a regularized cross-lingual visio-textual contrastive learning objective that constrains the representation proximity of weakly-aligned visio-textual inputs according to textual relevance. Besides, existing V&L pre-training approaches mainly deal with visual inputs by either region-of-interest (ROI) features or patch embeddings. We flexibly integrate the two forms of visual features into our model for pre-training and downstream multi-modal tasks. Extensive experiments on 5 downstream multi-modal tasks across 6 languages demonstrate the effectiveness of our proposed method over competitive contrast models with stronger zero-shot capability.

1 Introduction

Vision-language (V&L) pre-training aims to learn universal representations that can express visual and textual semantics informatively. It exploits a large amount of multi-modal data (*e.g.* image-text pairs) to make the model capable of handling cross-modal data. Till now, the advents of various
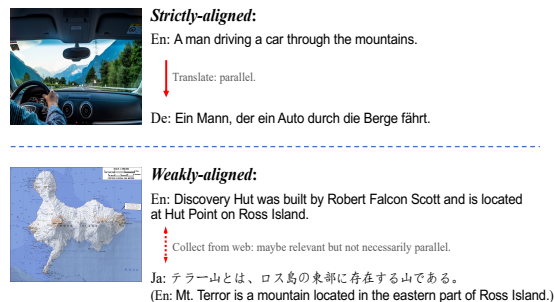

Figure 1: Comparison between "strictly-aligned" and "weakly-aligned" image-text pairs in different languages.

V&L pre-trained models have achieved remarkable results on many downstream multi-modal tasks.

Recently, V&L pre-trained models have developed from focusing on English-dominant tasks (Su et al., 2020; Chen et al., 2020; Cho et al., 2021) into multilingual scenarios (Ni et al., 2021; Liu et al., 2021a; Zhou et al., 2021). To this end, researchers construct multi-modal data in multiple languages and design various cross-lingual pre-training objectives. Such advances enable multi-modal modelling to leverage more diverse language resources. Meanwhile, these multilingual V&L pre-trained models also show their advantages over previous English-centric models in terms of generalization abilities across languages, especially in zero-shot settings.

Despite the promising performances of current multilingual V&L models, one of the major challenges is that they usually require massive strictly-aligned multilingual image-text pairs. The prevalent practice is to translate English-only multi-modal datasets into pseudo-parallel multilingual versions via machine translation (MT) (Ni et al., 2021; Zhou et al., 2021). However, the cost of collecting and translating such large-scale multi-modal datasets is often unbearable. To deal with this issue, we turn our eyes on those more easily available weakly-aligned multilingual multi-modal data, such as WIT (Srinivasan et al., 2021). As

shown in Figure 1, the so-called "weakly-aligned" means that the multilingual textual data of the same image are not strictly parallel.

In this paper, we propose a **R**egularized **C**ontrastive **C**ross-lingual **C**ross-modal (RC^3) pre-training framework, which can make better use of relatively abundant weakly-aligned multilingual image-text pairs. Specifically, we adopt an encoder-decoder architecture so that our model can be more adaptive to both discriminative and generative downstream tasks. Besides the widely used image-text matching (ITM) task, we further introduce masked conditional language modelling (MCLM) and cross-lingual textual contrastive learning (XTCL) along with our proposed regularized cross-lingual visio-textual contrastive learning (R-XVtCL) during pre-training. Particularly, while R-XVtCL encourages the visio-textual representations of two weakly-aligned image-text pairs to be close, a regularization term is designed to constrain such proximity according to the textual relevance of their respective texts.

Meanwhile, in current V&L models, there are mainly two ways of processing visual inputs:(1) Region-of-interest based (ROI-based). It uses external object detectors (*e.g.* Faster-RCNN (Ren et al., 2015b)) to extract ROI features from images and feed them with paired text into V&L models (Su et al., 2020; Chen et al., 2020; Cho et al., 2021; Ni et al., 2021; Liu et al., 2021a; Zhou et al., 2021). This method exerts the informativeness of ROI features, but such cumbersome protocol hinders the usage of massive online image-text pairs and requires additional procedures for various downstream tasks. (2) Patch-based. It directly transforms the original image pixels into patch embeddings and take them as inputs with textual data (Jia et al., 2021; Lee et al., 2022; Wang et al., 2022). This significantly simplifies pre-training protocols but cannot leverage informative ROI features. To improve the informativeness of visual features without complicating the whole training protocol, we flexibly integrate the above two forms of visual features into the model for pre-training and downstream tasks.

Our contributions can be summarized as follows: (1) We propose a cross-lingual cross-modal pre-training framework that can better exploit more abundant weakly-aligned multilingual image-text pairs; (2) We integrate ROI-based and patch-based visual features to enhance our V&L model for pre-training and downstream multi-modal tasks; (3) Ex-

tensive experiments on 5 downstream tasks across 6 languages show that our V&L model achieves higher or comparable performances over recent competitive contrast models with strong zero-shot capability.

2 Our Approach

In this section, we first briefly introduce the three types of datasets used for pre-training and more details are given in Appendix A. Then, we describe the model architecture and pre-training objectives.

2.1 Pre-training Data

Strictly-aligned Multilingual Image-caption Dataset D_s. We use the machine translation augmented image-caption paired data released in (Zhou et al., 2021). The English captions from Conceptual Captions dataset (Sharma et al., 2018) are translated into five different languages (Czech, German, French, Japanese and Chinese). This gives rise to a final strictly-aligned multilingual visio-linguistic dataset D_s, each image of which is paired with semantically-equivalent captions of 6 languages.

Weakly-aligned Multilingual Image-text Dataset D_w. We build a weakly-aligned visio-linguistic dataset D_w by extracting a fraction of multilingual image-caption pairs of 6 languages (German, English, French, Indonesian, Japanese and Chinese) from WIT dataset (Srinivasan et al., 2021). Note that the attached multilingual texts of the same image in D_w are not strictly parallel.

Multilingual Parallel Text Dataset D_t. We also use a combination of different textual data to form a multilingual parallel text dataset D_t. It is comprised of the parallel text corpus collected by (Zeng et al., 2022) from a subset of WikiMatrix (Schwenk et al., 2021a) and the parallel captions from D_s, which includes all 7 languages involved in D_s and D_w (*i.e.* English, Czech, German, French, Indonesian, Japanese and Chinese).

2.2 Model Architecture

We extend the encoder-decoder structure to make our model adaptive to both discriminative and generative multi-modal tasks. Figure 2 depicts the model architecture and the sequence formats for visio-textual/textual-only inputs.

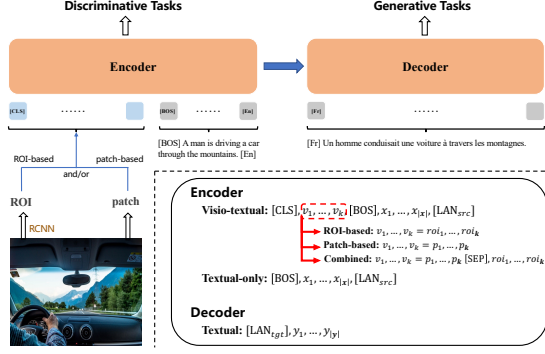**Cross-lingual Cross-modal Encoder.** As shown in Figure 2, given a visio-textual input composed

Figure 2: The architecture of our model and the sequence formats for visio-texual/textual-only inputs.

of an image and texts, the visual features are concatenated with text embeddings, which are then fed to the multi-layer encoder. Specifically, the visual features can be presented in the following three forms: (1) **ROI-based**. The ROI features roi = $\{roi_1, roi_2, ..., roi_k\}$ generated from an external object detector are projected by a fully-connected (FC) layer to have the same dimension as text embeddings; (2) **Patch-based**. Raw pixels are also mapped by another FC layer into a patch embedding sequence $p = \{p_1, p_2, ..., p_k\}$; (3) **Combined**. To enhance the informativeness of visual features, ROI features and patch embeddings are combined and fed to the encoder together, between which a special token [SEP] is inserted. For texts, we add a special token [BOS] and a language tag. Finally, a special token [CLS] is prepended at the beginning of the concatenated sequence, the output hidden state of which serves as its visio-textual representation (VtR).

For a textual-only input, only text embeddings are fed to the encoder and the output hidden state corresponding to [BOS] is used as its textual representation (TR).

Multilingual Decoder. In generative tasks that involve multiple languages, we also prepend a special language tag on the decoder side, indicating to which language the decoder is expected to generate texts.

2.3 Pre-training Objectives

During training, we adopt four pre-training tasks: (1) Masked Conditional Language Modelling (MCLM); (2) Image Text Matching (ITM); (3) Cross-lingual Textual Contrastive Learning (XTCL); (4) Regularized Cross-lingual Visio-textual Contrastive Learning (R-XVtCL). These tasks train the model to capture cross-lingual cross-

modal alignments among images and multilingual texts using different types of pre-training data described in Section 2.1.

2.3.1 Masked Conditional Language Modelling (MCLM)

Masked language modelling (MLM) has been widely used in previous encoder-only visio-linguistic models. Given an image v and its caption x^{l_i} in language l_i from the strictly-aligned dataset D_s, a word in x^{l_i} has a probability of 15% to be replaced with a special token [MASK]. The objective is to predict a set of masked words $x_m^{l_i}$ based on other unmasked words $x_{\backslash m}^{l_i}$ and the visual input:

$$L_{MLM} = -\mathbb{E}_{(v,x^{l_i})\sim D_s} \log P_{\theta_e}(x_m^{l_i}|x_{\backslash m}^{l_i}, v), \tag{1}$$

where θ_e is the trainable parameters of the encoder.

Moreover, with respect to x^{l_i}, since D_s also provides the parallel caption x^{l_j} in another language l_j, we simultaneously train the decoder to autoregressively predict the target text x^{l_j} based on the unmasked words $x_{\backslash m}^{l_i}$ and v. The MCLM objective can be formulated as follows:

$$L_{MCLM} = L_{MLM} \tag{2}$$

$$-\mathbb{E}_{(v,x^{l_i},x^{l_j})\sim D_s} \sum_{t=1}^{|x^{l_j}|} \log P_{\theta_d}(x_t^{l_j}|x_{<t}^{l_j}, x_{\backslash m}^{l_i}, v),$$

where θ_d is the trainable parameters of the decoder. In addition to MLM, the incorporation of the autoregressive term on the decoder can make the model better adapt to downstream generative tasks.

2.3.2 Image Text Matching (ITM)

ITM aims to discriminate whether an image and a piece of caption are matched, training the model to learn the alignment between visual and textual modalities. The representation of a visio-textual input (v, x^l) is fed to an FC layer and a sigmoid function to get a score $s_{\theta_e}(v, x^l)$. The score ranges from 0 to 1, predicting to what extent v and x^l are matched. We sample positive and negative visio-textual inputs from the strictly-aligned dataset D_s, where the negative one is constructed by randomly selecting another caption within the same batch to be paired with the original image. Thus, the training objective of ITM is written as

$$L_{ITM} = -\mathbb{E}_{(v,x^l)\sim D_s}[y \log s_{\theta_e}(v, x^l) \tag{3}$$
$$+(1-y) \log (1 - s_{\theta_e}(v, x^l))],$$

11749

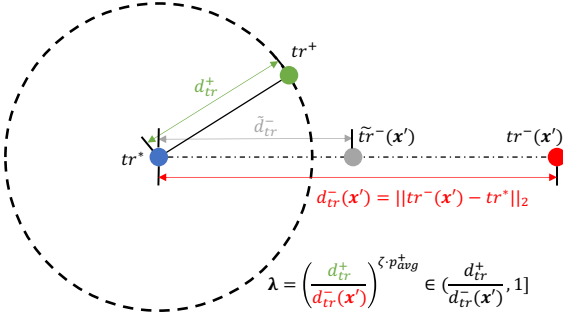**Universal Textual Representation Space (UTRS)**

Figure 3: The construction of harder negative samples by smoothed linear interpolation. The blue and green points represent the anchor instance and the positive sample in the UTRS, respectively. The red point refers to a negative sample whose interpolated harder TR representation corresponds to the grey one.

where $y \in \{0, 1\}$ indicates whether $(\boldsymbol{v}, \boldsymbol{x}^l)$ is a negative or positive sample.

2.3.3 Cross-lingual Textual Contrastive Learning (XTCL)

XTCL is to learn semantically informative representations of multilingual texts in a universal textual representation space (UTRS), where the TR representation of semantically equivalent texts are expected to be close while those of irrelevant ones are far from each other. Therefore, we adopt the interpolation-based contrastive learning method introduced in (Wei et al., 2021) to train the model, as depicted in Figure 3.

Specifically, given a batch of parallel text pairs $B_t = \{(\boldsymbol{x}_b^{l_i}, \boldsymbol{x}_b^{l_j})\}_{b=1}^{|B_t|}$ $(l_i \neq l_j)$ from the multilingual parallel dataset D_t, for a pair of parallel texts $(\boldsymbol{x}^{l_i}, \boldsymbol{x}^{l_j}) \in B_t$, we treat $\boldsymbol{x}^{l_i}$ as the anchor textual instance, the representation of which serves as the anchor point tr^* (the blue center) in the UTRS. Intuitively, the semantically equivalent $\boldsymbol{x}^{l_j}$ is naturally the positive sample and its representaion tr^+ (the green point on the circle) should be near to tr^*. On the contrary, each of the other texts $\boldsymbol{x}'$ within B is used as the negative sample whose TR representation $tr^-(\boldsymbol{x}')$, *i.e.* the red point out of the circle, should be far from the anchor. The XTCL objective can be defined as

$$L_{xltcl}(\boldsymbol{x}^{l_i}) = -\log \frac{\exp{(-d_{tr}^+)}}{\exp{(-d_{tr}^+)} + \sum\limits_{\boldsymbol{x}' \in \mathcal{N}(\boldsymbol{x}^{l_i})} \exp{(-d_{tr}^-(\boldsymbol{x}'))}}, \tag{4}$$

where $\mathcal{N}(\boldsymbol{x}^{l_i})$ is the set of negative samples with respect to $\boldsymbol{x}^{l_i}$, d_{tr}^+ and $d_{tr}^-(\boldsymbol{x})$ denote the euclidean TR distances from tr^+ and each $tr(\boldsymbol{x}'^-)$ to the anchor in the UTRS, *i.e.* $d_{tr}^+ = ||tr^+ - tr^*||_2$ and

$d_{tr}^-(\boldsymbol{x}') = ||tr^-(\boldsymbol{x}') - tr^*||_2.$

However, since the above negative samples are usually not informative, following (Wei et al., 2021), we generate harder negative samples by smoothed linear interpolation (Bowman et al., 2016; Zheng et al., 2019). For a negative sample $\boldsymbol{x}'$ from N_{tr}, a more difficult negative representation in the UTRS is constructed through the following interpolation:

$$\widetilde{tr}^-(\boldsymbol{x}') = \begin{cases} tr^* + \boldsymbol{\lambda}(tr^-(\boldsymbol{x}') - tr^*), & d_{tr}^-(\boldsymbol{x}') > d_{tr}^+; \\ tr^-(\boldsymbol{x}'), & d_{tr}^-(\boldsymbol{x}') \leq d_{tr}^+; \end{cases} \tag{5}$$

$$\boldsymbol{\lambda} = \left(\frac{d_{tr}^+}{d_{tr}^-(\boldsymbol{x}')}\right)^{\zeta \cdot p_{avg}^+}, \tag{6}$$

where $p_{avg}^+ = \frac{1}{100}\sum_{\tau \in [-100,-1]} e^{-L_{xltcl}^{(\tau)}}$ is the average log-probability over the previous 100 training steps in Equation 4 and ζ is a slacking coefficient set to 0.9 in our experiment. By doing so, the difficulty of the interpolated representation $\widetilde{tr}^-(\boldsymbol{x}')$, *i.e.* the grey point out of the circle in Figure 3, can be dynamically adjusted during training, which results in a lower $\boldsymbol{\lambda}$ (harder samples) when p_{avg}^+ increases and vice versa.

Thus, Equation 4 is reformulated by replacing the original representation of each negative sample $\boldsymbol{x}'$ with the harder interpolated one $\widetilde{tr}^-(\boldsymbol{x}')$:

$$\tilde{L}_{xltcl}(\boldsymbol{x}^{l_i}) = -\log \frac{\exp{(-d_{tr}^+)}}{\exp{(-d_{tr}^+)} + \sum\limits_{\boldsymbol{x}' \in \mathcal{N}(\boldsymbol{x}^{l_i})} \exp{(-\tilde{d}_{tr}^-(\boldsymbol{x}'))}}, \tag{7}$$

where $\tilde{d}_{tr}^-(\boldsymbol{x}')$ is the euclidean distance between the anchor and $\widetilde{tr}^-(\boldsymbol{x}')$, *i.e.* $\tilde{d}_{tr}^-(\boldsymbol{x}') = ||\widetilde{tr}^-(\boldsymbol{x}') - tr^*||_2$. Finally, the XTCL objective is:

$$L_{XTCL} = \mathbb{E}_{(\boldsymbol{x}^{l_i}, \boldsymbol{x}^{l_j}) \sim D_t} \tilde{L}_{xltcl}(\boldsymbol{x}^{l_i}). \tag{8}$$

In this way, the relevance of two arbitrary pieces of texts can be measured by the proximity of their TR representations in the UTRS, which will be used in the next pre-training objective.

2.3.4 Regularized Cross-lingual Visio-textual Contrastive Learning (R-XVtCL)

Similarly, the R-XVtCL objective is to learn semantically informative representations of visio-textual inputs in a universal visio-textual representation space (UVtRS), which involves both strictly-aligned and weakly-aligned image-caption pairs. We treat visio-textual inputs in another representation space because they differ from textual-only inputs in that their semantics depend on both images
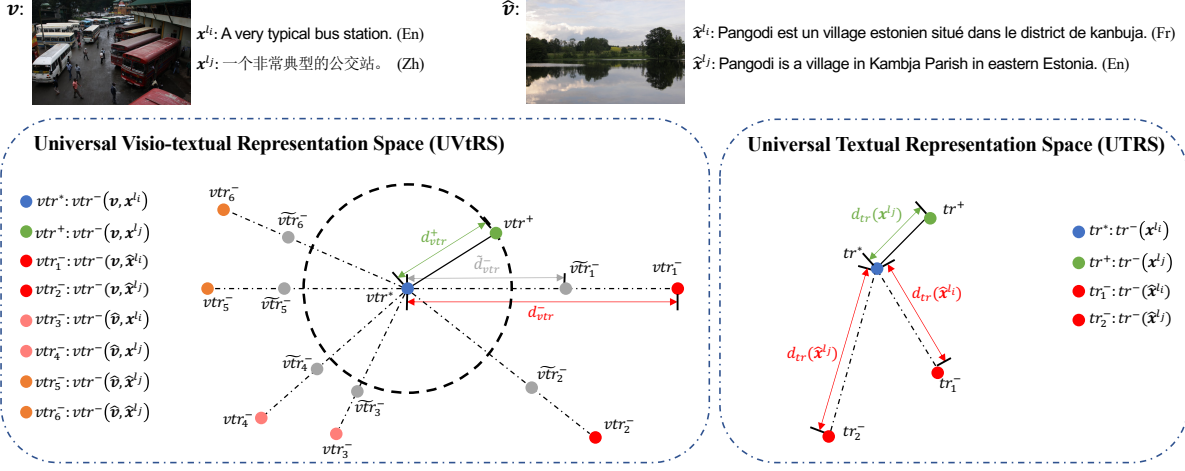
Figure 4: The illustration of Regularized Cross-lingual Visio-textual Contrastive Learning (R-XVtCL). In the UVtRS, the red, pink and brown points out of the circle correspond to VtR representations of the three types of negative samples. The grey ones are the interpolated VtR representation that are harder for the model to discriminate. In the UTRS, the distance between any two TR representations measures the textual relevance of their related texts.

and texts. Analogously, we also expect the visio-textual representations (VtR) of semantically equivalent visio-textual inputs are near to each other.

First, we introduce how to leverage the strictly-aligned multilingual image-caption pairs. Given a batch of image-caption triplets in two different languages $B_{vt}=\{(\boldsymbol{v}_b, \boldsymbol{x}_b^{l_i}, \boldsymbol{x}_b^{l_j})\}_{b=1}^{|B_{vt}|}$ ($l_i \neq l_j$), for a triplet $(\boldsymbol{v}, \boldsymbol{x}^{l_i}, \boldsymbol{x}^{l_j}) \in B_{vt}$, we use the pair $(\boldsymbol{v}, \boldsymbol{x}^{l_i})$ as the anchor visio-textual instance, with its VtR representation vtr^* serving as the anchor point in the UVtRS. Meanwhile, since $\boldsymbol{x}^{l_j}$ is parallel to $\boldsymbol{x}^{l_i}$, the pair $(\boldsymbol{v}, \boldsymbol{x}^{l_j})$ is used as the positive sample, whose VtR representation vtr^+ should be close to vtr^*. Along with $(\boldsymbol{v}, \boldsymbol{x}^{l_i}, \boldsymbol{x}^{l_j})$, we construct three types of negative visio-textual samples using another triplet $(\hat{\boldsymbol{v}}, \hat{\boldsymbol{x}}^{l_i}, \hat{\boldsymbol{x}}^{l_j})$ within the same batch:

(1) $(\boldsymbol{v}, \hat{\boldsymbol{x}}^{l_i})$ and $(\boldsymbol{v}, \hat{\boldsymbol{x}}^{l_j})$, containing the same image with the anchor instance but semantically non-equivalent captions;

(2) $(\hat{\boldsymbol{v}}, \boldsymbol{x}^{l_i})$ and $(\hat{\boldsymbol{v}}, \boldsymbol{x}^{l_j})$, containing semantically equivalent captions but different paired images;

(3) $(\hat{\boldsymbol{v}}, \hat{\boldsymbol{x}}^{l_i})$ and $(\hat{\boldsymbol{v}}, \hat{\boldsymbol{x}}^{l_j})$, containing different images and semantically non-equivalent captions.

With these negative samples, we construct their harder representations in the UVtRS through the similar interpolation procedure described in Section 2.3.3, resulting in their interpolated VtR representations, as illustrated in Figure 4.[1] Therefore,

---

[1] We denote these VtR representations as $\widetilde{vtr}^-(\boldsymbol{v}, \hat{\boldsymbol{x}}^{l_i})$,

the contrastive loss using strictly-aligned multilingual image-caption pairs can be written as

$$\tilde{L}_{xlvtcl}(\boldsymbol{v}, \boldsymbol{x}^{l_i}) = \tag{9}$$
$$-\log \frac{\exp(-d_{vtr}^+)}{\exp(-d_{vtr}^+) + \sum\limits_{(\boldsymbol{v}', \boldsymbol{x}') \in \mathcal{N}(\boldsymbol{v}, \boldsymbol{x}^{l_i})} \exp(-\tilde{d}_{vtr}^-(\boldsymbol{v}', \boldsymbol{x}'))},$$

where $\mathcal{N}(\boldsymbol{v}, \boldsymbol{x}^{l_i})$ includes the above three types of negative samples, d_{vtr}^+ and $\widetilde{d}_{vtr}^-(\boldsymbol{v}', \boldsymbol{x}')$ are the euclidean distances from vtr^+ and each interpolated $\widetilde{vtr}^-(\boldsymbol{v}', \boldsymbol{x}')$ to the anchor in the UVtRS.

However, when using weakly-aligned multilingual image-caption pairs, it is not reasonable to simply encourage the VtR representation vtr^+ to be close to the anchor vtr^* because $\boldsymbol{x}^{l_i}$ and $\boldsymbol{x}^{l_j}$ are not strictly parallel. Hence, we propose to constrain the proximity of $(\boldsymbol{v}, \boldsymbol{x}^{l_j})$ to the anchor instance $(\boldsymbol{v}, \boldsymbol{x}^{l_i})$ in the UVtRS through an additional regularization term, given that the proximity of two TR representations in the UTRS can be seen as textual relevance (See Section 2.3.3).

Concretely, we first obtain the TR representations of all captions in the two weakly-aligned image-caption triplets $(\boldsymbol{v}, \boldsymbol{x}^{l_i}, \boldsymbol{x}^{l_j})$ and $(\hat{\boldsymbol{v}}, \hat{\boldsymbol{x}}^{l_i}, \hat{\boldsymbol{x}}^{l_j})$ from D_w. The textual relevances of $\boldsymbol{x}^{l_j}$, $\hat{\boldsymbol{x}}^{l_i}$ and $\hat{\boldsymbol{x}}^{l_j}$ with respect to $\boldsymbol{x}^{l_i}$ can be measured by the negative TR distance, i.e. $-d_{tr}(\boldsymbol{x}^{l_j})$, $-d_{tr}(\hat{\boldsymbol{x}}^{l_i})$ and $-d_{tr}(\hat{\boldsymbol{x}}^{l_j})$, the closer to 0 the more relevant. Then, we transform these relevance scores into a normal-

$\widetilde{vtr}^-(\boldsymbol{v}, \hat{\boldsymbol{x}}^{l_j})$, $\widetilde{vtr}^-(\hat{\boldsymbol{v}}, \boldsymbol{x}^{l_i})$, $\widetilde{vtr}^-(\hat{\boldsymbol{v}}, \boldsymbol{x}^{l_j})$, $\widetilde{vtr}^-(\hat{\boldsymbol{v}}, \hat{\boldsymbol{x}}^{l_i})$ and $\widetilde{vtr}^-(\hat{\boldsymbol{v}}, \hat{\boldsymbol{x}}^{l_j})$.

ized relevance distribution in the UTRS:

$$P_{tr} = \text{softmax}([-d_{tr}(\boldsymbol{x}^{l_j}), -d_{tr}(\hat{\boldsymbol{x}}^{l_i}), -d_{tr}(\hat{\boldsymbol{x}}^{l_j})]).$$
(10)

Moreover, in the UVtRS, we can also obtain such a normalized relevance distribution P_{vtr}. Concretely, we select the image-text pairs that contain the same image as the anchor visio-textual instance $(\boldsymbol{v}, \boldsymbol{x}^{l_i})$, including $(\boldsymbol{v}, \boldsymbol{x}^{l_j})$, $(\boldsymbol{v}, \hat{\boldsymbol{x}}^{l_i})$ and $(\boldsymbol{v}, \hat{\boldsymbol{x}}^{l_j})$, because their VtR representation differences with the anchor only derive from semantically non-equivalent texts. Thereafter, P_{vtr} can be computed as

$$P_{vtr} = \text{softmax}([-d_{vtr}(\boldsymbol{v}, \boldsymbol{x}^{l_j}),$$
$$- d_{vtr}(\boldsymbol{v}, \hat{\boldsymbol{x}}^{l_i}), -d_{vtr}(\boldsymbol{v}, \hat{\boldsymbol{x}}^{l_j})]).$$
(11)

Hence, the regularized contrastive loss with weakly-aligned multilingual image-text pairs is:

$$\tilde{L}_{xlvtcl}^{reg}(\boldsymbol{v}, \boldsymbol{x}^{l_i}) = \tilde{L}_{xlvtcl}(\boldsymbol{v}, \boldsymbol{x}^{l_i}) + \text{KL}(P_{vtr}||P_{tr}).$$
(12)

Finally, with training instances from both D_s and D_w, the R-XVtCL objective can be formulated as the following:

$$L_{R-XVtCL} = \mathbb{E}_{(\boldsymbol{v}, \boldsymbol{x}^{l_i}, \boldsymbol{x}^{l_j}) \sim D_s} \tilde{L}_{xlvtcl}(\boldsymbol{v}, \boldsymbol{x}^{l_i}) +$$
$$\mathbb{E}_{(\boldsymbol{v}, \boldsymbol{x}^{l_i}, \boldsymbol{x}^{l_j}) \sim D_w} \tilde{L}_{xlvtcl}^{reg}(\boldsymbol{v}, \boldsymbol{x}^{l_i}). \quad (13)$$

Note that D_s and D_w are simultaneously used in this task. In particular, images from D_s are processed into ROI-based visual features while those from D_w are in the form of patch-based features. This is due to the fact that in general scenarios, the cost of obtaining ROI features of all images from much more abundant weakly-aligned image-text data is often unbearable.

3 Experiments

3.1 Downstream Tasks

We conduct experiments on five downstream multi-modal tasks across 6 languages (English, German, French, Indonesian, Japanese and Chinese), including Cross-lingual Visual Natural Language Inference (**XVNLI**), Cross-lingual Grounded Question Answering (**xGQA**), Multicultural Reasoning over Vision and Language (**MaRVL**), Image-Text Retrieval (**ITR**) and Multi-modal Machine Translation (**MMT**). The first four are discriminative tasks while the last one is a generative task. The details about these tasks and their datasets are given in Appendix B.

Model/Task	XVNLI	xGQA	MaRVL	ITR
M^3P	76.89	53.75	68.22	27.97
mUNITER	76.38	54.68	**71.91**	**42.70**
xUNITER	75.77	54.83	71.55	35.25
UC^2	76.38	55.19	70.56	35.97
RC^3-Patch	71.21	41.36	-	-
RC^3-ROI	77.91	54.13	69.42	41.12
RC^3-Combined	**78.43**	**55.92**	69.74	41.30

Table 1: Performances on English testsets of XVNLI, xGQA, MaRVL and ITR tasks. We report the average scores under three different random seeds.

3.2 Implementation Details

Following the setting of MBart-50 (Tang et al., 2020), our model consists of 12 encoder layers and 12 decoder layers with 16 attention heads and 1024 hidden dimensions, which is initialized by MBart-50 parameters. For visual inputs, the dimension of ROI-based features and patch embeddings are 2048 and 768, respectively. We use the ROI features provided in IGLUE (Bugliarello et al., 2022) generated from Faster-RCNN (Ren et al., 2015a), which contain 36 regions for each image. Every original image is resized to 224×224 pixels and then mapped to a flattened one-dimensional patch embedding sequence, where the patch size is set to 32×32. For text inputs, we build a vocabulary out of the original one used in MBart-50, achieving a cover rate of over 99.99% on the seven languages involved in our pre-training and downstream tasks.

During pre-training, we use Adam optimizer (Kingma and Ba, 2015) with a learning rate of 5×10^{-5}. We use DeepSpeed to support multi-node training. It takes about ten days to converge on 64 V100 GPUs, where the model is updated for 100,000 steps and the batch size is set to 1024. More details are given in Appendix B.

3.3 Contrast Models

For the four discriminative tasks, we compare our model with recent competitive multilingual V&L pre-trained models trained with strictly-aligned multilingual image-caption dataset: **M³P** (Ni et al., 2021), **mUNITER**, **xUNITER** (Liu et al., 2021a) and **UC²** (Zhou et al., 2021). Meanwhile, we make comparison with several strong baselines, including **MeMAD** (Grönroos et al., 2018), **VL-T5** and **VL-BART** (Cho et al., 2021). All of these contrast models leverage ROI-based visual features during their pre-training and fine-tuning.

Model/Task	XVNLI	xGQA			MaRVL		ITR		
	Fr	De	Id	Zh	Id	Zh	De	Ja	Zh
M^3P	56.36	33.42	32.58	28.65	56.47	55.04	12.60	9.95	15.60
mUNITER	59.36	23.95	9.36	7.03	54.79	55.34	11.95	7.00	11.60
xUNITER	63.32	34.83	**33.73**	19.55	55.14	53.06	13.95	10.50	15.87
UC^2	69.67	42.85	28.67	31.16	56.74	59.88	26.25	23.32	28.95
RC^3-Patch	64.43	24.44	22.53	25.97	-	-	-	-	-
RC^3-ROI	71.65	40.39	29.24	36.06	**57.80**	**62.55**	**35.20**	30.82	35.52
RC^3-Combined	**72.43**	**43.69**	31.94	**39.49**	57.26	60.77	34.35	**30.90**	**37.20**

Table 2: Zero-shot performances on non-English XVNLI, xGQA, MaRVL and ITR testsets. We also report the average scores under three different random seeds.

3.4 Evaluation on Discriminative Tasks

In our experiments, we fine-tune the pre-trained model using only the English training data of each task and evaluate its performance on each target language, which means that the evaluations on non-English languages follow a zero-shot setting. The metrics of XVNLI, xGQA and MaRVL are accuracy and that of ITR is Recall@1. Note that there are two retrieval directions in ITR task: image-to-text and text-to-image, where the average Recall@1 on the two directions is reported in Table 1 and Table 2. We denote our V&L model trained using Patch-based, ROI-based and Combined visual features as **RC^3-Patch**, **RC^3-ROI** and **RC^3-Combined**, respectively. The reported results of other contrast models are provided in IGLUE benchmark (Bugliarello et al., 2022).

Results on English Testsets. From Table 1, we can observe that RC^3-Combined achieves better results on the English testsets of XVNLI and xGQA tasks over other contrast models, slightly underperforming mUNITER, xUNITER and UC^2 on MaRVL. Meanwhile, the ITR results of RC^3-ROI and RC^3-Combined surpass all other models except the best performing mUNITER. Another phenomenon is that the inferiority of RC^3-Patch over other variants indicates the importance of informativeness from visual features on these tasks, especially MaRVL and ITR where RC^3-Patch is uncomparably worse. Whereas, RC^3-Combined performs better than RC^3-ROI, showing that additional patch embeddings still benefit the model to some extent.

Zero-shot Results. Table 2 gives the zero-shot performances on XVNLI, xGQA, MaRVL and ITR tasks across multiple non-English languages. Overall, we can see that our models, RC^3-ROI and RC^3-Combined, significantly outperform other contrast

Model/Testset	En-De		En-Fr	
	2016	2017	2016	2017
MeMAD	38.9	32.0	62.2	54.4
VL-T5	45.5	40.9	-	-
VL-BART	41.3	35.9	-	-
RC^3-Patch	45.49	**42.06**	68.29	62.56
RC^3-ROI	45.73	41.52	68.38	**62.71**
RC^3-Combined	**45.86**	42.01	**68.50**	62.66

Table 3: Performances on Multi30k English-to-German (En-De) and English-to-French (En-Fr) testsets.

models. Particularly for ITR, the zero-shot results of our models exceed the strongest UC^2 model by considerable margins in all three languages. As for xGQA, though M^3P and xUNITER perform slightly better in Indonesian, our model RC^3-Combined still achieves higher accuracy in German (43.69 v.s 42.85) and especially Chinese (39.49 v.s 31.16).

For MaRVL, it can be seen that although RC^3-Combined surpasses other contrast models, it is inferior to RC^3-ROI. We conjecture that this is due to the double-image nature of MaRVL task.[2] Concretely, when "Combined" visual features of the two involved images are fed together to the encoder, the excessive length of visual inputs might distract the model from adequately attending to the textual modality, which cannot offset the benefit gained from additional patch embeddings. Such effect particularly stands out in a zero-shot setting, where V&L models more heavily rely on meaningful textual representations learned from pre-training and the English-only fine-tuning.

3.5 Evaluation on MMT

MMT is a generative task that involves both encoder and decoder to generate translations based on source sentences and their paired images. Table 3 lists the performances on Mulit30K English-

---

[2]Please refer to Appendix B for the details about MaRVL task and its specific visual input formats.

Model/Task	XVNLI		xGQA				MaRVL			ITR			
	En	Fr	En	De	Id	Zh	En	Id	Zh	En	De	Ja	Zh
RC^3-Combined	**78.43**	**72.43**	**55.92**	43.69	**31.94**	39.49	**69.74**	57.26	60.77	**41.30**	34.35	**30.90**	**37.25**
RC^3-ROI	77.91	71.65	54.13	40.39	29.24	36.06	69.42	**57.80**	**62.55**	41.12	**35.22**	30.82	35.55
w/o. $\mathrm{KL}(P_{vtr}||P_{tr})$	76.26	70.69	53.41	43.49	24.32	33.27	69.41	55.31	58.89	40.60	34.72	29.15	35.20
w/o. R-XVtCL	74.34	70.26	52.63	**44.87**	20.06	**41.93**	69.28	50.62	57.41	40.82	34.02	30.72	35.02
w/o. R-XVtCL & XTCL	76.17	70.52	52.43	39.71	11.17	33.28	68.83	52.21	56.42	39.22	33.80	30.02	34.70

Table 4: Ablation results. Note that all variants except RC^3-Combined adopt ROI-based visual features for evaluation.

to-German (En-De) and English-to-French (En-Fr) datasets. We can see that our models outperform other contrast models. Nevertheless, according to previous research (Caglayan et al., 2019), it shows that the source sentences in Multi30k dataset presumably take more effects than images for translations, which could explain the outcome that our three model variants exhibit no obvious differences.

4 Ablation Study

In this section, we conduct ablation studies to investigate the effect of our proposed training objectives in Section 2.3. Adopting ROI-based visual features, we investigate the following three model variants:

- *w/o.* $\mathrm{KL}(P_{vtr}||P_{tr})$: This variant removes the regularization term in Equation 12, which means that the weakly-aligned multilingual image-caption pairs from D_w are used in the same way as strictly-aligned ones.

- *w/o. R-XVtCL*: In this variant, the R-XVtCL objective is totally removed during pre-training.

- *w/o. R-XVtCL & XTCL*: In this variant, we remove both XTCL and R-XVtCL objectives, only using MCLM and ITM for pre-training.

From Table 4, it is clear that the removal of $\mathrm{KL}(P_{vtr}||P_{tr})$ in Equation 12 gives rise to performance drops, which demonstrates the effectiveness of constraining the VtR representation proximity of multilingual weakly-aligned image-caption pairs. In Appendix C, we give several illustrative cases that present how our proposed textual relevance-based regularization affects the VtR representation proximity in the UVtRS. Moreover, although *w/o. R-XVtCL* achieves the highest accuracy on German and Chinese xGQA datasets, it still mostly underperforms compared to *w/o.* $\mathrm{KL}(P_{vtr}||P_{tr})$, RC^3-ROI and RC^3-Combined. This shows that the R-XVtCL objective brings improvement to our model by enhancing the learned VtR representations. Besides, removing both R-XVtCL and

XTCL results in worse performances compared to the other two ablation variants except on XVNLI.

5 Related Work

In recent years, there have been a series of V&L pre-trained models achieving remarkable progress on many downstream multi-modal tasks. Overall, these studies adjust model architectures and design various pre-training objectives to learn alignment between visual and textual modalities. They can be mainly classified into single-stream (Chen et al., 2020; Cho et al., 2021; Wang et al., 2022) and two-stream V&L architectures (Lu et al., 2019; Zeng et al., 2022).

Apart from the above models, some multilingual V&L pre-trained models are proposed to learn universal representations across multiple languages and modalities. One of the major difficulties is the lack of high-quality multilingual multi-modal pre-training data. To address this issue, Ni et al. (2021) proposed to integrate multilingual pre-training and multi-modal pre-training. Concretely, batches of multilingual text corpora and monolingual multi-modal data are alternately used. Following a similar manner, Liu et al. (2021a) build *mUNITER* and *xUNITER* by initializing model parameters with *mBERT* and *XLM-R*, respectively. Furthermore, Zhou et al. (2021) translate original English pre-training data into multiple languages and propose UC^2 to learn universal representations by introducing two cross-lingual/cross-modal pre-training tasks. These models leverage strictly-aligned multilingual and multi-modal datasets that are relatively difficult to collect. Therefore in this paper, we additionally make better use of more abundant weakly-aligned multilingual multi-modal data.

6 Conclusion

In this paper, we propose Regularized Contrastive Cross-lingual Cross-modal pre-training, which additionally exploits relatively more abundant weakly-aligned multilingual image-text pairs. During pre-training, we constrain the proximity of

visio-textual representations of weakly-aligned image-text pairs according to their textual relevance. Besides, we further enhance our V&L model by integrating ROI-based and patch-based visual features. Compared with recent competitive V&L models, our model achieves higher or comparable results, especially demonstrating stronger zero-shot performance.

Limitations

Currently, we build a vocabulary from the original one used in MBart-50, and only conduct downstream experiments across 6 languages (English, German, French, Indonesian, Japanese and Chinese). Although we could involve more languages, it would require a larger CUDA memory that might go beyond our device capacity. Hence, we merely select the above languages that have sufficient overlap with our pre-training datasets. In addition, for fair comparisons, we only use the strictly-aligned multilingual multi-modal dataset provided in (Zhou et al., 2021), which is augmented through machine translation. It is unclear how the quality of strictly-aligned dataset would affect model performance. Meanwhile, the length of texts in our weakly-aligned multilingual multi-modal dataset is generally very long. As a result, we truncate textual inputs before feeding them into the encoder, possibly bringing information loss to some extent.

References

Željko Agić and Natalie Schluter. 2018. Baselines and test data for cross-lingual inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL*, pages 10–21.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and

Ivan Vulic. 2022. IGLUE: A benchmark for transfer learning across modalities, tasks, and languages. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2370–2392.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4159–4170.

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 *Computer Vision - ECCV 2020 - 16th European Conference*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 1931–1942.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Mérialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphaël Troncy, and Raúl Vázquez. 2018. The memad submission to the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018*, pages 603–611.

Estevam Hruschka, Tom Mitchell, Dunja Mladenic, Marko Grobelnik, and Nikita Bhutani, editors. 2022. *Proceedings of the 2nd Workshop on Deriving Insights from User-Generated Text*. Association for Computational Linguistics, (Hybrid) Dublin, Ireland, and Virtual.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*.

Youhan Lee, Kyungtae Lim, Woonhyuk Baek, Byungseok Roh, and Saehoon Kim. 2022. Efficient multilingual multi-modal pre-training through triple

contrastive loss. In *Proceedings of the 29th International Conference on Computational Linguistics*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021a. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021b. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems*, pages 13–23.

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. 2021. M3P: learning universal representations via multitask multilingual multimodal pretraining. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3977–3986.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublin, Ireland. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015a. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of NIPS*, volume 28.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015b. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems*, pages 91–99.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. 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*, pages 1351–1361.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021b. 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*, pages 1351–1361, Online. Association for Computational Linguistics.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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*, pages 2556–2565.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: wikipedia-based image text dataset for multimodal multilingual machine learning. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449. ACM.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pretraining of generic visual-linguistic representations. In *8th International Conference on Learning Representations,*.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022. Simvlm: Simple visual language model pretraining with weak supervision. In *The Tenth International Conference on Learning Representations*.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *CoRR*, abs/1901.06706.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hock-enmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78.

Yan Zeng, Wangchunshu Zhou, Ao Luo, and Xinsong Zhang. 2022. Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training. *CoRR*, abs/2206.00621.

Wenzhao Zheng, Zhaodong Chen, Jiwen Lu, and Jie Zhou. 2019. Hardness-aware deep metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 72–81. Computer Vision Foundation / IEEE.

Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. 2021. UC2: universal cross-lingual cross-modal vision-and-language pre-training. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4155–4165.

A Pre-training Data

As described in Section 2.1, our pre-training involves three types of data:

Strictly-aligned Multilingual Image-caption Dataset D_s. Following previous work (Bugliarello et al., 2022), we use the ConceptCaption dataset as the strictly-aligned multilingual image-caption dataset D_s, which contains the original 2,777,649 image-caption pairs and machine-translated captions in five other languages (Czech, German, French, Japanese and Chinese). Besides, during pre-training, we use the pre-processed ROI features provided in IGLUE benchmark (Bugliarello et al., 2022)

Weakly-aligned Multilingual Image-text Dataset D_w. This dataset is built from a fraction of the publicly-available WIT (Hruschka et al., 2022) dataset. In WIT, there are a large number of unique images that have multiple pieces of related texts in different languages. First, we index images through their unique urls. Then, each image is paired with multiple pieces of related texts of different languages, resulting in a multilingual image-text tuple $(v, x^{l_i}, x^{l_j}, ..., x^{l_k})$ that shares the same image. The statistics of the constructed weakly-aligned dataset is provided in Table 5, where each entry represents the number of multilingual image-text tuples in the corresponding language pair.

Multilingual Parallel Text Dataset D_t. For this dataset used in XTCL task, we combine the parallel texts from D_s and a subset of WikiMatrix (Schwenk et al., 2021b) used in (Zeng et al., 2022). As a result, D_t contains multilingual parallel texts of 7 languages, covering all languages involved in the pre-training and all downstream tasks, *i.e.* Czech, German, French, Indonesian, Japanese and Chinese.

B Downstream Tasks and Datasets

We conduct experiments on five downstream multi-modal tasks: XVNLI, xGQA, MaRVL, ITR and MMT. For all downstream tasks, we fine-tune the model on English training sets, and then evaluate performances across all languages. The hyperparameters used in our experiments are listed in Table 6.

XVNLI. Cross-lingual Visual Natural Language Inference task aims to discriminate whether a given textual hypothesis *entails*, *contradicts*, or is *neutral*

an image premise. Its dataset combines three existing text-only datasets SNLI (Bowman et al., 2015), with their cross-lingual (Agić and Schluter, 2018) and multi-modal (Xie et al., 2019) counterparts.

xGQA. The goal of Cross-lingual Grounded Question Answering task is to answer several types of structured questions about an image. The corresponding dataset is manually translated from the GQA (Pfeiffer et al., 2022) validation set into 7 languages.

MaRVL. Multicultural Reasoning over Vision and Language task (Liu et al., 2021b) requires the model to determine whether a textual description is true or false about a pair of images. Following (Bugliarello et al., 2022), the NLVR2 dataset (Suhr et al., 2019) is used for training while the MaRVL dataset is used for testing. Because the V&L model needs to take in two images as inputs in this task, the input format of visual features is different from other tasks. Specifically, given a piece of text x and an image pair (v^1, v^2), we concatenate visual and textual features as [CLS], $v_1^1, v_2^1, ..., v_k^1$, [SEP′], $v_1^2, v_2^2, ..., v_k^2$, [BOS], x_1, $x_2, x_{|x|}$, [LAN$_{src}$], where a special token [SEP′] is inserted between two images. In the same way, the top-layer hidden state corresponding to [CLS] is used as the final visio-textual representation for fine-tuning and evaluation.

ITR. Image-Text Retrieval task is composed of image-to-text and text-to-image retrieval. Image-to-text retrieval is to select out the most relevant texts from a candidate set given an image. Inversely, text-to-image retrieval is to pick the most relevant image. We also use the ITR dataset provided in (Bugliarello et al., 2022), which is collected by combining 1,000 images from Flickr30K (Young et al., 2014) and 1,000 from MSCOCO (Lin et al., 2014).

MMT. Multi-modal Machine Translation task is to translate a source sentence with the help of its paired image. We conduct experiments on the widely-used Multi30k dataset (Elliott et al., 2016), where each image is paired with one English description and human translations into German&French. The training and validation sets contain 29,000 and 1,014 instances, respectively. Besides, the test sets consist of *test2016* and *test2017*, each of which contains 1,000 instances for evaluation.

	En	De	Fr	Ja	Zh	Id
En	5,157,134	739,697	814,485	376,759	357,677	163,442
De	739,697	3,248,830	516,048	199,996	163,226	77,632
Fr	814,485	516,048	2485,944	223,177	188,968	91,712
Ja	376,759	199,996	223,177	1,032,183	174,226	67,030
Zh	357,677	163,226	188,968	174,226	798,853	66,294
Id	163,442	77,632	91,712	67,030	66,294	266,144

Table 5: Detailed statistics of weakly-aligned multilingual image-text dataset D_w.

(a) 1. $\boldsymbol{x}^{En}$: Crown jewels are the objects of metalwork and jewelry in the regalia of a current or former monarchy…

2. $\boldsymbol{x}^{De}$: Die Tiara, Papstkrone oder auch gelegentlich römische Krone genannt, ist die früher bei feierlichen Anlässen getragene Krone des Papstes ...
(Translation: The tiara, pope's crown or occasionally Roman crown, is the crown of the pope, formerly worn on solemn occasions…)

(b) 3. $\boldsymbol{x}^{En}$: It is used with various meats, seafood and vegetables in stews, soups, barbecue, sotos, gulai, and also as an addition to Indonesian-style instant noodles …

4. $\boldsymbol{x}^{Fr}$:Le nasi campur est un plat de nasi putih accompagné d'autres plats en petites portions, tels que de la viande, des légumes, des arachides, des œufs et des...
(Translation: Nasi Camp is a dish of Nasi Putih, accompanied by other small dishes, such as meat, vegetables, peanuts, eggs and vegetables ...)

(c) 5. $\boldsymbol{x}^{En}$: Novarossi World, also known as Novarossi Nitro Micro Engines, are an Italian manufacturer of model engines and related items for radio-controlled models …

6. $\boldsymbol{x}^{Ja}$:ノ ヴァロッシ ワールドはノヴァロッシの商標のグローエンジンとラジコン用の関連する製品群で有名なイタリアの模型用小型エンジンの会社である…
(Translation: Nova Rossi world is a small engine model for Italy's model engine, famous for Nova Rossi's trademark glow engine and related products for radios...)

(d) 7. $\boldsymbol{x}^{En}$: The 1999 Pacific typhoon season was the last Pacific and it ran year-round in 1999...

8. $\boldsymbol{x}^{Zh}$: 1999年太平洋台风季泛指在1999年全年内的任何时间….
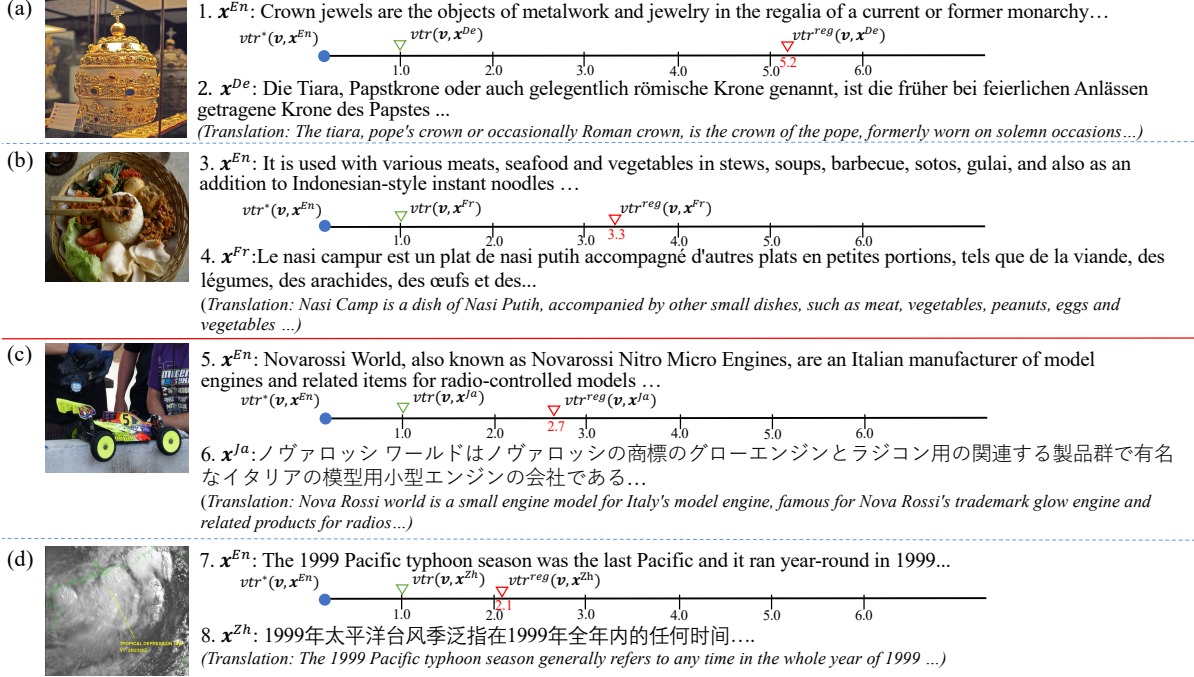*(Translation: The 1999 Pacific typhoon season generally refers to any time in the whole year of 1999 ...)*

Figure 5: Illustrative cases. For the axis of each case, the blue start point represents the anchor VtR representation. $vtr^*(\boldsymbol{v}, \boldsymbol{x}^{En})$. The green positions on the axis represent the ratio unit 1.0, corresponding to the VtR representation without being regularized with $\text{KL}(P_{vtr}||P_{tr})$. The red positions refer to the regularized VtR representation in terms of distance ratio in the UVtRS.

Hyperparameters	XVNLI	xGQA	MaRVL
Learning Rate	4e-5	4e-5	4e-5
Batch size	128	256	64
Epochs	10	5	40
Input length	80	40	80
Hyperparameters	ITR	MMT (En-De)	MMT (En-Fr)
Learning Rate	1e-5	5e-6	5e-6
Batch size	64	256	256
Epochs	10	5	5
Input length	80	50	50

Table 6: Hyperparameters for downstream tasks.

C   Case Study

In Figure 5, we exhibit several typical cases that can show the effect of our proposed regularization term $\text{KL}(P_{vtr}||P_{tr})$ in Equation 12, each of which contains an image and two pieces of texts. For each case, the image and its English texts are combined as the anchor visio-textual instance $vtr^*(\boldsymbol{v}, \boldsymbol{x}^{En})$, corresponding to the blue start point in Figure 5.

Similarly, the combination of the image and its non-English texts serves as the target visio-textual input whose euclidean VtR distance from $vtr^*(\boldsymbol{v}, \boldsymbol{x}^{En})$ is worth probing. We introduce an axis to indicate the proximity of non-English visio-textual input to the anchor in the UVtRS with and without $\text{KL}(P_{vtr}||P_{tr})$.

Taking (a) for instance, let $vtr(\boldsymbol{v}, \boldsymbol{x}^{De})$ and $vtr^{reg}(\boldsymbol{v}, \boldsymbol{x}^{De})$ represent the VtR representations with and without regularization, respectively. We compute their euclidean distances to the anchor, denoted as d_{vtr} and d_{vtr}^{reg}. Instead of marking the two absolute distances on the axis, we choose to record their ratio d_{vtr}^{reg}/d_{vtr} that can reflect the proximity change after adding the regularization term $\text{KL}(P_{vtr}||P_{tr})$. This is because the relative proximity is the what really matters for each case. Referring to the translations in italics, we can observe that the paired texts in cases (c) and (d) are more

relevant to each other, *i.e.* 1↔2 and 3↔4, than those in (a) and (b), *i.e.* 5↔6 and 7↔8. Accordingly, it is clearly shown that the proximity changes of VtR representations are more significant in cases (a) and (b).

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section "Limitation".

☐ A2. Did you discuss any potential risks of your work?
Not applicable. No risk.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section "Abstract" and Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
No.

B ☑ Did you use or create scientific artifacts?

Section 3.

☑ B1. Did you cite the creators of artifacts you used?
Section 3.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 3.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 3.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. We just used data released by previous researchers.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix.

C ☑ Did you run computational experiments?

Section 3.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 3.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3 and Appendix.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

No.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Not applicable.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Not applicable.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Not applicable.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Not applicable.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Not applicable.