X^2-VLM: All-In-One Pre-trained Model For Vision-Language Tasks

Yan Zeng[*] ByteDance Research
Xinsong Zhang ByteDance Research
Hang Li ByteDance Research
Jiawei Wang ByteDance Research
Jipeng Zhang HKUST
Wangchunshu Zhou ETH Zurich

Abstract

Vision language pre-training aims to learn alignments between vision and language from a large amount of data. Most existing methods only learn image-text alignments. Some others utilize pre-trained object detectors to leverage vision language alignments at the object level. In this paper, we propose to learn multi-grained vision language alignments by a unified pre-training framework that learns multi-grained aligning and multi-grained localization simultaneously. Based on it, we present X^2-VLM, an all-in-one model with a flexible modular architecture, in which we further unify image-text pre-training and video-text pre-training in one model. X^2-VLM is able to learn unlimited visual concepts associated with diverse text descriptions. Experiment results show that X^2-VLM performs the best on base and large scale for both image-text and video-text tasks, making a good trade-off between performance and model scale. Moreover, we show that the modular design of X^2-VLM results in high transferability for it to be utilized in any language or domain. For example, by simply replacing the text encoder with XLM-R, X^2-VLM outperforms state-of-the-art multilingual multi-modal pre-trained models without any multilingual pre-training. The code and pre-trained models are available at `github.com/zengyan-97/X2-VLM`.

1 Introduction

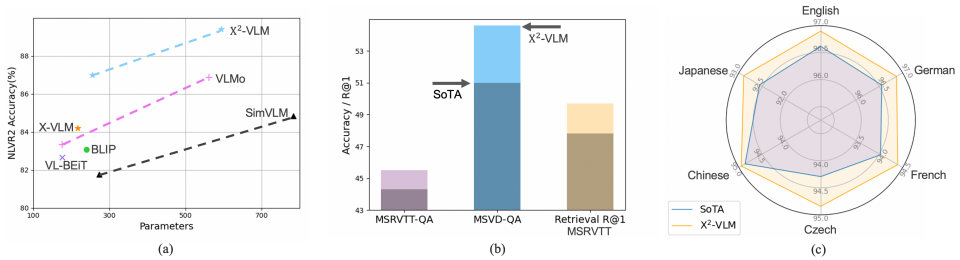


Figure 1: (a) Comparison of X^2-VLM with existing image-text pre-training methods on the visual reasoning task. (b) Comparison with existing video-text pre-training methods on video-text tasks. (c) Comparison with existing multilingual multi-modal pre-training methods.

Vision language pre-training aims to learn vision language alignments from a large number of image-text or video-text pairs. A pre-trained Vision Language Model (VLM) fine-tuned with a small amount

[*]Correspondence to: <zengyan.yanne@bytedance.com>.

of labeled data has shown state-of-the-art (SoTA) performances in many Vision Language (V+L) tasks, such as image-text retrieval and visual question answering (VQA).

Existing work learning vision language alignments typically falls into two categories: *coarse-grained* and *fine-grained*. Coarse-grained approaches use convolutional neural networks [1] or vision transformers [2] to encode overall image features [3–5], which however have difficulties in learning fine-grained vision language alignments, e.g., at the object level, from noisy image-text pairs which are usually weak-correlated [6]. To learn fine-grained vision language alignments, many approaches adopt pre-trained object detectors as the image encoder [7–13]. However, object detectors output object-centric features unable to encode relations among multiple objects. Moreover, an object detector can only recognize a limited number of object categories.

Ideally, a VLM should simultaneously learn multi-grained alignments between vision and language in pre-training without being restricted to object-text alignments or image-text alignments. However, learning multi-grained alignments is challenging, and previous work has failed to handle this issue. The challenges come from four aspects: 1) what types of data to use to learn multi-grained vision language alignments; 2) how to aggregate the different types of data in a unified way for vision language pre-training; 3) how to represent multi-grained visual concepts, including objects, regions, and images, by a single vision encoder; 4) how to efficiently learn multi-grained vision language alignments from the data.

In this paper, we present an all-in-one VLM pre-trained with a unified framework to learn multi-grained vision language alignments, namely X^2-VLM. We leverage three types of data for vision language pre-training, including object labels on images [14–16] such as "man" or "backpack", region annotations on images [17, 16] such as "boy wearing backpack", and text descriptions for images such as "The first day of school gives a mixed feeling to both students and parents.". We assume that learning multi-grained vision language alignments can help VLMs better understand weak-correlated image-text pairs since the model has learned to align the components in images, e.g., objects or regions, to textual descriptions, e.g., words or phrases. We associate all visual concepts with text descriptions instead of class labels, including objects, regions, and images. By associating all visual concepts with language, the model can learn unlimited visual concepts described by diverse texts in a unified way.

X^2-VLM has a flexible modular architecture, with three modules for vision, text, and fusion, respectively. All modules are based on Transformer [18]. We encode an image with vision transformer [2], and we utilize certain patch features to represent multi-grained visual concepts in the image that can be objects, regions, or the image itself. By doing so, X^2-VLM outputs vision features for objects, regions, and images in a unified form. Furthermore, we propose directly aligning the multi-grained vision features with the paired text features and simultaneously locating multi-grained visual concepts in the same image given different text descriptions for vision language pre-training. In fine-tuning and inference, X^2-VLM can leverage the learned multi-grained alignments to perform the downstream V+L tasks without object or region annotations in the input images.

X^2-VLM can be easily extended to video-text pre-training. For video encoding, we sample video frames and encode the frames with vision transformer respectively. Then, we use the average in the temporal dimension of patch features of frames to encode a video. The encoder parameters are shared between video-text pre-training and image-text pre-training. By doing so, we leverage video-text pairs to enable the model to understand visual concepts in temporal dimension and learn a more versatile VLM.

Moreover, we show the flexibility of X^2-VLM with the modular architecture. We investigate whether the cross-modal ability can be transferred to other languages or domains after pre-training. This is an important problem in real-world applications because many multi-modal tasks exist in non-English languages. However, since collecting image-text pairs or video-text pairs in certain languages or domains can be costly, recent SoTA VLMs are trained with English data and only applicable to English texts, limiting their application scopes. We find that surprisingly X^2-VLM can effectively adapt to V+L tasks in different languages or domains by simply replacing the text module with a language-specific or domain-specific one without further pre-training.

We conduct extensive experiments to verify the effectiveness of X^2-VLM. First, we compare X^2-VLM with SoTA image-text pre-training methods on base and large scale and find that X^2-VLM substantially outperforms all of them in the image-text tasks, including retrieval, VQA, reasoning, and

grounding. Moreover, X^2-VLM outperforms SimVLM [19] and BLIP [20], which are designed for generative tasks, in image caption generation. X^2-VLM also outperforms MDETR [21] and OFA [22], which also leverage image annotations of objects and regions, in cross-modal understanding tasks. X^2-VLM$_{\text{large}}$ with $\sim$590M parameters performs competitively to CoCa [23] and BEiT-3 [24] with $\sim$2B parameters, especially on image-text retrieval and visual reasoning. In summary, X^2-VLM makes a good trade-off between performance and model scale, as indicated in Figure 1 (a). Besides, we find that by training with large-scale image-text pairs, X^2-VLM learns to locate diverse fine-grained visual concepts in open-domain images, such as different sodas, cars, characters and celebrities. Second, X^2-VLM is also the new SoTA pre-trained model on video-text tasks, including video-text retrieval and video VQA, as shown in Figure 1 (b). Most existing VLMs only tackle image-text tasks, but X^2-VLM with a unified framework achieves SoTA performances on both types of tasks. Third, to verify the flexibility of the modular design, we replace the text encoder of X^2-VLM with XLM-R [25], a multilingual text encoder, after vision-language pre-training on English data. As indicated in Figure 1 (c), X^2-VLM outperforms SoTA multilingual multi-modal pre-training methods that need multilingual image-text pairs [26, 27] and multilingual sentence pairs [28] which are costly to collect.

The contributions of this paper are as follows:

- We propose to learn multi-grained vision language alignments by a unified pre-training framework that learns multi-grained aligning and multi-grained localization simultaneously. Based on it, we present X^2-VLM, an all-in-one pre-trained VLM that can handle both image-text and video-text tasks.
- Experiment results show that X^2-VLM is the best model on base and large scale on both image-text and video-text benchmarks. Furthermore, the results confirm that the proposed framework for multi-grained vision language pre-training is scalable to massive data and larger model sizes.
- We reveal the potential of the modular design of X^2-VLM, showing that it can be utilized in other languages or domains. By replacing the text encoder with XLM-R after pre-training on English data, X^2-VLM outperforms SoTA methods on multi-lingual multi-modal tasks.

2 Related Work

2.1 Image-Text Pre-training

The existing work on image-text pre-training typically falls into two categories: fine-grained and coarse-grained. Fine-grained approaches [7–13] utilize a pre-trained object detector [29, 30] as the image encoder, which is trained on annotations of common objects, e.g. COCO [14] and Visual Genome [17]. An object detector first identifies all regions that probably contain an object, then conducts object classification on each region. An image is then represented by dozens of object-centric features of the identified regions. However, object-centric features cannot represent relations among multiple objects in different regions. Therefore, it is difficult for this approach to effectively encode multi-grained visual concepts. Moreover, object detectors can only detect common objects, e.g. only 80 object categories for the COCO dataset. Thus, it is suboptimal to apply this approach to encode various visual concepts in real-world applications. For example, the approach cannot distinguish "Pepsi" from "Coca Cola" or "Audi" from "BMW".

In contrast, the coarse-grained approaches build VLMs by extracting and encoding overall image features with convolutional network [31, 3] or vision transformer [4, 5]. While being more efficient, the performance of the coarse-grained approach is usually not as good as the fine-grained approach since the latter leverages vision language alignments at the object level, which are shown to be critical for downstream V+L tasks. However, with advanced vision transformers, e.g. Swin-Transformer [32] and BEiT-2 [33], recent methods such as METER [34] and VL-BEiT [35], can outperform strongest fine-grained approach VinVL [13].

There also emerge some methods attempting to learn both object-level and image-level alignments. However, these approaches still rely on object detectors and thus suffer from the aforementioned problems. For example, E2E-VLP [36] adds an end-to-end object detection module (i.e. DETR [37]). KD-VLP [38] relies on external object detectors to perform object knowledge distillation. Different

from these approaches, our framework for multi-grained vision language pre-training does not rely on object detection, and it learns vision language alignments not restricted to object-level or image-level in a unified way.

2.2 Video-Text Pre-training

Most existing VLMs only tackle image-text tasks. Only a few VLMs work on video-text pre-training. Since a video consists of multiple images, video-text models usually share many similarities with image-text models in both model architecture and training objectives. Though video-text pre-training shares similarities with image-text pre-training, no existing method can achieve SoTA performances on both types of tasks. Representative work on video-text pre-training including ClipBERT [39], Frozen [40], ALPRO [41], VIOLET [42], and All-in-one [43]. There are other methods optimized specifically for a downstream task, for either video-text retrieval [44, 45] or video question answering [46]. Recently, OmniVL [47] is proposed to support both image-text tasks and video-text tasks. It utilizes 3D patch embeddings for videos and 2D patch embeddings for images, and adopts TimeSformer [48] for vision encoding.

2.3 Multilingual Multi-modal Pre-training

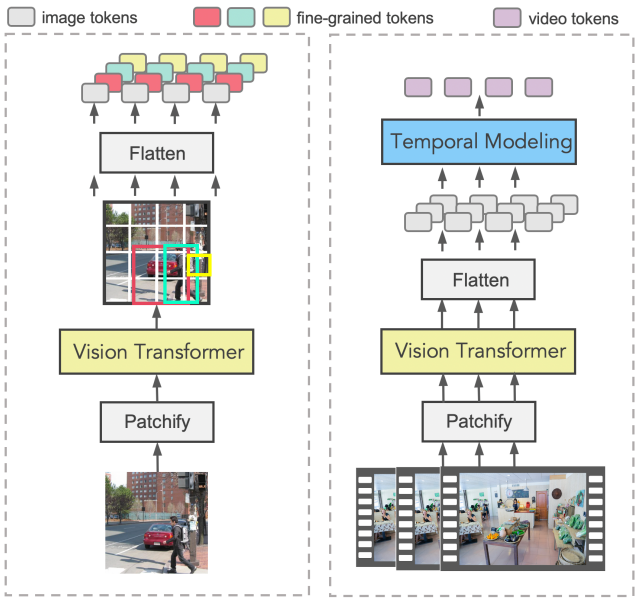
Multilingual multi-modal pre-training aims to make multi-modal models applicable to non-English texts. While appealing, multi-lingual multi-modal pre-training has its own challenges. Unlike multi-lingual pre-training and multi-modal pre-training where a relatively large amount of parallel data is available, there exist only a few multi-lingual multi-modal corpora and their language coverage is also limited. Therefore, M^{3}P [49] utilizes 101G texts covering 100 languages for pre-training. It makes English a pivot and alternates between English-only vision-language pre-training and multi-lingual masked language modeling. Differently, UC^2 [26] translates image-text pairs in English into five different languages and uses all the data for pre-training. MURAL [27] collects large-scale image-text pairs in 110 languages. CCLM [28] utilizes parallel multilingual text pairs and proposes a simple framework that unifies cross-lingual and cross-modal pre-training with shared architecture and objectives. All these methods require extra data to perform multilingual multi-modal pre-training. In contrast, we show that X^2-VLM can adapt to multilingual V+L tasks without the need for multilingual multi-modal pre-training process by exploiting the potential of its modular architecture.

3 Method

3.1 Overview

Architecture: X^2-VLM consists of vision, text, and multi-modal fusion modules. All modules are based on Transformer [18]. The fusion module takes text features as input and fuses the vision features with the text features through cross-attention at each layer, where the text features work as the queries and the vision features work as the keys and values. In pre-training, the three modules work as encoders, while the text and fusion modules can also be adapted for generation tasks if applying left-to-right self-attention as shown in our experiments for image caption generation. Figure 3 illustrates the architecture of X^2-VLM and the way we perform multi-grained aligning and multi-grained localization.

Data: X^2-VLM is a unified approach that associates all visual concepts with text descriptions, including image-text pairs, video-text pairs, and image annotations of objects and regions. That is to say, an image may contain more than one visual concept and each of them is associated with a text description, denoted as $(I, T, \{(V^1, T^1), ...\}^N)$. $\{(V^1, T^1), (V^2, T^2)...\}^N$ are the image annotations of objects or regions. Here, V^i is an object or region in a bounding box $\mathbf{b}^i = (cx, cy, w, h)$ represented by the normalized center coordinates, width, and height of the box. When the image itself represents a visual concept, $\mathbf{b} = (0.5, 0.5, 1, 1)$. T^i for objects are originally object labels. If an object annotation contains object attributes, e.g. color, we concatenate the attribute with the object label as the text description. T^i for regions are phrases that describe the regions. Note that, as listed in Table 1, some images do not have associated texts, i.e., T is NaN, and some images do not have annotations, i.e., $N = 0$. **Nevertheless, we mix all types of data in a training batch, and thus for each training iteration, we optimize the model by multi-grained aligning and multi-grained localization simultaneously.**



(a) Image Encoding (a) Video Encoding

Figure 2: **Unified vision encoding** For images, we extract the subset of patch features from the vision transformer to represent an image and objects/regions in the image. For videos, each frame is first encoded independently, and than a light-weight non-parametric temporal modeling layer is applied across frames.

3.2 Unified Vision Encoding

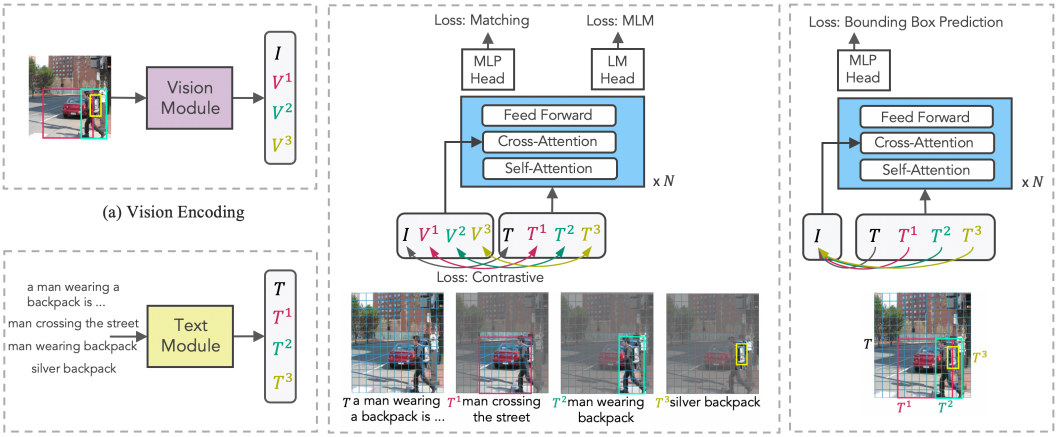
X^2-VLM unifies image and video encoding, as illustrated in Figure 2. Irrespective of the inputs, the vision module of X^2-VLM produces hidden states in the latent feature space of the vision transformer. As a result, image-text pre-training and fine-grained pre-training mutually reinforce one another. Moreover, the capability of image understanding can be better transferred for video comprehension.

Visual Concept Representation X^2-VLM proposes an efficient way to obtain all multi-grained visual concepts in an image with only one forward pass of the vision transformer. First, we process an image into patch features. Then, X^2-VLM represents an object or a region, e.g. V^i, that corresponds to a set of patches in the bounding box, e.g. $\mathbf{b}^i$, by aggregating information among the patches as illustrated in Figure 2. Specifically, we flatten the corresponding patch features while keeping their original positions. Then, we calculate the average of the patch features as the [CLS] patch and prepend it. Accordingly, the representation of the entire image I is obtained by aggregating information among all the patches.

Video Representation Since a video consists of multiple images, to leverage large-scale image-text pre-training for better video understanding, we unify video encoding and image encoding in a simple and efficient way. First, we sample one frame per second for videos. Then, for each training iteration, we randomly sample a few frames of a video. The vision encoder will encode the frames into patch features respectively. Finally, we add temporal information to the patch features of each frame and calculate the average in temporal dimension to represent the video. By doing so, a video is encoded by a sequence of patch features the same as an object/region/image, and thus we can apply a unified pre-training framework for both video-text pairs and object/region/image-text pairs.

3.3 Multi-Grained Vision Language Pre-training

We mix all types of data in a training batch, and thus for each training iteration, as shown in Figure 3, we optimize X^2-VLM by two objectives simultaneously: 1) learning multi-grained alignments between visual concepts and texts; 2) locating multi-grained visual concepts in images given different text descriptions.



(a) Vision Encoding (b) Text Encoding (c) Multi-Grained Aligning (d) Multi-Grained Localization

Figure 3: **Illustration of the proposed multi-grained vision language pre-training.** X^2-VLM consists of vision, text, and fusion modules. After encoding visual concepts (a) and text inputs (b), multi-grained vision features are then paired with corresponding text features for multi-grained aligning (c). Besides, the image is paired with different textual descriptions for multi-grained localization to predict the bounding box for each visual concept (d). All the datasets we used are publicly available (see Section 4.1).

3.3.1 Multi-Grained Aligning

Since we have associated all visual concepts with text descriptions, we propose to align the multi-grained visual concepts with the corresponding texts. Specifically, after encoding visual concepts by the aforementioned method, we align the vision features in multiple granularities with the corresponding text features in the same way. We simply choose three losses for optimization, including contrastive loss, matching loss, and MLM loss. These losses have been well-studied by previous work [11, 50, 5], but we propose to employ them on the visual concept-to-text level. Note that V in this section represents a visual concept, including an object, region, image, or video.

We apply contrastive loss to predict (visual concept, text) pairs from in-batch negatives. Given a pair (V, T), T is the positive example for V, and we treat the other $(N-1)$ texts within the mini-batch as negative examples. First, we define the similarity by:

$$s(V, T) = g_v(\mathbf{v}_{\text{cls}})^\top g_w(\mathbf{w}_{\text{cls}}), \tag{1}$$

where $\mathbf{v}_{\text{cls}}$ and $\mathbf{w}_{\text{cls}}$ are the output `[CLS]` embedding of the vision encoder and the text encoder respectively. g_v and g_w are transformations that map the `[CLS]` embeddings to normalized lower-dimensional representations. Based on it, we calculate the in-batch vision-to-text similarity as:

$$p^{\text{v2t}}(V) = \frac{\exp(s(V, T)/\tau)}{\sum_{i=1}^{N} \exp(s(V, T^i)/\tau)}, \tag{2}$$

Similarly, the text-to-vision similarity is:

$$p^{\text{t2v}}(T) = \frac{\exp(s(V, T)/\tau)}{\sum_{i=1}^{N} \exp(s(V^i, T)/\tau)}, \tag{3}$$

where τ is a learnable temperature parameter. Let $\mathbf{y}^{\text{v2t}}(V)$ and $\mathbf{y}^{\text{t2v}}(T)$ denote the ground-truth one-hot similarity, in which only the positive pair has the probability of one. Finally, the contrastive loss is defined as the cross-entropy H between $\mathbf{p}$ and $\mathbf{y}$:

$$\mathcal{L}_{\text{cl}} = \frac{1}{2}\mathbb{E}_{V,T\sim D}\left[\text{H}(\mathbf{y}^{\text{v2t}}(V), \mathbf{p}^{\text{v2t}}(V)) + \text{H}(\mathbf{y}^{\text{t2v}}(T), \mathbf{p}^{\text{t2v}}(T))\right] \tag{4}$$

We also utilize the matching loss to determine whether a pair of visual concept and text is matched. For each visual concept in a mini-batch, we sample an in-batch hard negative text by following $p^{\text{v2t}}(V)$ in Equation 2. Texts that are more relevant to the concept are more likely to be sampled. We

also sample one hard negative visual concept for each text. We then put the pairs as inputs for the fusion module, and then we use $\mathbf{x}_{\text{cls}}$, the output `[CLS]` embedding of the fusion module, to predict the matching probability p^{match}, and the loss is:

$$\mathcal{L}_{\text{match}} = \mathbb{E}_{V,T\sim D}\text{H}(\mathbf{y}^{\text{match}}, \mathbf{p}^{\text{match}}(V,T)), \tag{5}$$

where $\mathbf{y}^{\text{match}}$ is a 2-dimensional one-hot vector representing the ground-truth label.

Furthermore, we apply masked language modeling loss to predict the masked words in the text based on the visual concept. We randomly mask out the input tokens with a probability of 40%, and the replacements are 10% random tokens, 10% unchanged, and 80% `[MASK]`. We use the fusion encoder's outputs and append a linear layer followed by softmax for prediction. Let $\hat{T}$ denote a masked text, and $\mathbf{p}^j(V,\hat{T})$ denote the probability of the masked token t_j predicted by the fusion module. We minimize the cross-entropy loss:

$$\mathcal{L}_{\text{mlm}} = \mathbb{E}_{t_j\sim T;(V,T)\sim D}\text{H}(\mathbf{y}^j, \mathbf{p}^j(V,\hat{T})), \tag{6}$$

where $\mathbf{y}^j$ is a one-hot distribution in which the ground-truth token t_j has the probability of one.

3.3.2 Multi-Grained Localization

We have aligned visual concepts with texts in different granularity. We further optimize X^2-VLM by training it to locate different visual concepts in the same image given corresponding text descriptions. Specifically, we introduce bounding box prediction task into vision language pre-training, where the model is asked to predict the bounding box $\mathbf{b}^i = (cx, cy, w, h)$ of a visual concept V^i:

$$\hat{\mathbf{b}}^i(I,T^i) = \text{Sigmoid}(\text{MLP}(\mathbf{x}^i_{\text{cls}})), \tag{7}$$

where Sigmoid is for normalization, MLP denotes multi-layer perceptron, and $\mathbf{x}^i_{\text{cls}}$ is the output `[CLS]` embedding of the fusion module given the features of I (the entire image) and T^i (the description of the visual concept).

For bounding box prediction, ℓ_1 is the most commonly-used loss. However, it has different scales for small and large boxes, even if their relative errors are similar. To mitigate this issue, we use a linear combination of the ℓ_1 loss and the generalized Intersection over Union (IoU) loss [51], which is scale-invariant. The overall loss is defined as:

$$\mathcal{L}_{\text{bbox}} = \mathbb{E}_{(V^i,T^i)\sim I;I\sim D}[\mathcal{L}_{\text{iou}}(\mathbf{b}^i, \hat{\mathbf{b}}^i) + ||\mathbf{b}^i - \hat{\mathbf{b}}^i||_1] \tag{8}$$

Finally, the pre-training objective of X^2-VLM is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{bbox}} + \mathcal{L}_{\text{cl}} + \mathcal{L}_{\text{match}} + \mathcal{L}_{\text{mlm}} \tag{9}$$

4 Experiment

4.1 Pre-training Datasets

We pre-train X^2-VLM with two sets of data. The 4M pre-training dataset consists of two in-domain datasets, COCO [14] and Visual Genome (VG) [17], and two out-of-domain datasets, SBU Captions [52] and Conceptual Captions (CC) [53]. This pre-training dataset is widely utilized by previous work, and thus we choose this setting to make a fair comparison with other methods. We also include annotations for COCO and VG images from RefCOCO [54], GQA [55], and Flickr entities [56] following OFA [22] and MDETR [21].

Then, we scale up the pre-training dataset by including out-of-domain and much noisier image-text pairs from Conceptual 12M dataset (CC-12M) [57] and LAION [58], and object annotations from Objects365 [15] and OpenImages [16]. Besides, to support video-text downstream tasks, we include video-text pairs from WebVid2.5M [40], Howto100M [59], and YT-Temporal 180M [60] for pre-training. Note that all the datasets we used are public available and have been exploited in previous work [5, 13, 20, 22, 43]. Besides, since most downstream tasks are built on top of COCO and VG, we exclude all images that also appear in the test sets of downstream tasks to avoid information leak. We give data filtering details in Appendix.

Dataset	# Images	# Captions	# Objects	# Regions
COCO	0.11M	0.55M	0.45M	-
VG	0.10M	-	2.0M	3.7M
SBU	0.86M	0.86M	-	-
CC-3M	2.9M	2.9M	-	-
CC-12M	11.1M	11.1M	-	-
Objects365	0.58M	-	2.0M	-
OpenImages	1.7M	-	4.2M	-
LAION	1.3B	1.3B	-	-
WebVid2.5M	2.5M	2.5M	-	-
Howto100M	1.7M	1.7M	-	-
YTT180M	5.3M	5.3M	-	-

Table 1: **Statistics of the pre-training datasets.** We pre-train X^2-VLM with two sets of data: one contains COCO, VG, SBU, and CC-3M, where the total number of images is 4M; the other one includes more noisy image-text pairs and video-text pairs.

Model	Hidden	Vision		Text		Fusion	
		Layers	Params	Layers	Params	Layers	Params
X^2-VLM$_{\text{base}}$	768	12	86M	12	111M	6	55M
X^2-VLM$_{\text{large}}$	1024	24	303M	12	190M	6	95M

Table 2: **Size variants of X^2-VLM.** All modules consist of transformer layers.

4.2 Implementation Details

Table 2 lists the parameters of X^2-VLM. Considering the trade-off between performance and model scale [61], X^2-VLM$_{\text{large}}$ also uses a 12L text encoder. The vision encoder is initialized with BEiT-2 [33]. The text encoder is initialized with BERT [62]. X^2-VLM is pre-trained at image resolution of 224×224 using 16×16 patch size. We mix all types of data in a training batch, and thus for each training iteration, we optimize the model by multi-grained aligning and multi-grained localization simultaneously. With 4M data, we pre-train X^2-VLM$_{\text{base}}$ for 500K steps with a batch size of 1024 on 8 A100 and X^2-VLM$_{\text{large}}$ for 250K steps on 16 A100, which takes ~ 1 week. The learning rate of X^2-VLM$_{\text{base}}$ is warmed-up to $1e^{-4}$ in the first 2500 steps and decayed following a linear schedule. The learning rate is $5e^{-5}$ for X^2-VLM$_{\text{large}}$. With large-scale data, training X^2-VLM takes 2-3 weeks on 32 A100 for the base model and 64 A100 for the large model. We describe the implementation details in Appendix.

4.3 Image-Text Downstream Tasks

We compare X^2-VLM with the most well-known state-of-the-art approaches on five widely used image-text downstream tasks. In general, we follow the settings in the previous work on fine-tuning. We describe how we implement fine-tuning as follows.

4.3.1 Image-Text Retrieval

We evaluate X^2-VLM on both MSCOCO and Flickr30K [56] datasets. We adopt the widely used Karpathy split [63] for both datasets. We optimize $\mathcal{L}_{\text{cl}}$ and $\mathcal{L}_{\text{match}}$ for fine-tuning. We set the batch size to 1024. The resolution of input images is set to 384x384. Following the previous work [5], X^2-VLM first encodes images and texts separately and calculates in-batch text-to-image and image-to-text similarities to obtain the top-k candidates, and then uses the fusion encoder to re-rank the candidates. k is set to 80 for the MSCOCO dataset and 32 for Flickr30K.

Table 3 shows that X^2-VLM achieves SoTA results on image-text retrieval tasks especially on Flickr30K benchmark even though existing approaches either have more model parameters or more

Model	# Params	MSCOCO (5K test set)						Flickr30K (1K test set)					
		TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Models Pretrained on COCO, VG, SBU and CC datasets (4M)													
ALBEF	210M	73.1	91.4	96.0	56.8	81.5	89.2	94.3	99.4	99.8	82.8	96.7	98.4
$\text{VLMo}_{\text{base}}$†	175M	74.8	93.1	96.9	57.2	82.6	89.8	92.3	99.4	99.9	79.3	95.7	97.8
VL-BEiT	175M	79.5	-	-	61.5	-	-	95.8	-	-	83.9	-	-
OmniVL	288M	76.8	93.6	97.3	58.5	82.6	89.5	94.9	99.6	99.9	83.4	97.0	98.6
$\text{X}^2\text{-VLM}_{\text{base}}$	255M	**80.5**	**95.5**	**97.8**	**62.7**	**84.7**	**90.7**	**97.4**	**99.9**	**100**	**90.0**	**98.6**	**99.3**
$\text{VLMo}_{\text{large}}$†	562M	78.2	94.4	97.4	60.6	84.4	91.0	95.3	99.9	100	84.5	97.3	98.6
$\text{X}^2\text{-VLM}_{\text{large}}$	593M	**82.3**	**96.2**	**98.3**	**65.2**	**86.4**	**91.9**	**99.1**	**100**	**100**	**91.1**	**98.6**	**99.4**
Models Pretrained on More Data													
$\text{BLIP}_{\text{base}}$	240M	81.9	95.4	97.8	64.3	85.7	91.5	97.3	99.9	100	87.3	97.6	98.9
OmniVL	288M	82.1	95.9	98.1	64.8	86.1	91.6	97.3	99.9	100	87.9	97.8	99.1
$\text{X}^2\text{-VLM}_{\text{base}}$	255M	**83.5**	**96.3**	**98.5**	**66.2**	**87.1**	**92.2**	**98.5**	**100**	**100**	**90.4**	**98.2**	**99.3**
ALIGN†	490M	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	100	84.9	97.4	98.6
FLIP†	420M	78.9	94.4	97.4	61.2	84.3	90.6	96.6	100	100	87.1	97.7	99.1
$\text{BLIP}_{\text{large}}$	452M	82.4	95.4	97.9	65.1	86.3	91.8	97.4	99.8	99.9	87.6	97.7	99.0
$\text{X}^2\text{-VLM}_{\text{large}}$	593M	**84.4**	**96.5**	**98.5**	**67.7**	**87.5**	**92.5**	**98.8**	**100**	**100**	**91.8**	**98.6**	**99.5**

Table 3: **Results of image-to-text retrieval (TR) and text-to-image retrieval (IR) on COCO and Flickr30K.** † denotes dual-encoder retrieval models, and others use a fusion module to re-rank top-k candidates following ALBEF [5].

	# Params	MSCOCO		Flickr30K	
		TR	IR	TR	IR
BEiT-3	1.9B	**84.8**	67.2	98.0	90.3
$\text{X}^2\text{-VLM}_{\text{large}}$	593M	84.4	**67.7**	**98.8**	**91.8**

Table 4: X^2-VLM compared with the SoTA giant model, BEiT-3, on image-text retrieval benchmarks. We report Recall@1 for both image-to-text retrieval (TR) and text-to-image retrieval (IR).

training data. Concretely, $\text{X}^2\text{-VLM}_{\text{base}}$ outperforms FLIP [64], $\text{BLIP}_{\text{base}}$ and $\text{BLIP}_{\text{large}}$ which also exploits large-scale image-text pairs from LAION, and $\text{X}^2\text{-VLM}_{\text{large}}$ further improves the image-text retrieval performance. Compared to OmniVL which also supports both image-text tasks and video-text tasks, $\text{X}^2\text{-VLM}_{\text{base}}$ substantially outperforms it when pre-trained with the 4M data or with more data. These results validate the advantage of learning multi-grained vision language alignments.

We also compare $\text{X}^2\text{-VLM}_{\text{large}}$ with BEiT-3, a giant foundation model with 1.9B model parameters in Table 4. Experimental results show that though being much smaller, $\text{X}^2\text{-VLM}_{\text{large}}$ has a comparable or even better performance compared with BEiT-3. Moreover, as shown in Table 3, $\text{X}^2\text{-VLM}_{\text{base}}$ substantially outperforms VL-BEiT which is the base version of BEiT-3 in the 4M setting. On the other hand, when comparing X^2-VLM's performances in different settings in Table 3, we can see that the proposed framework for multi-grained vision language pre-training has good scalability which can benefit from a larger model size and large-scale out-of-domain image-text pairs.

4.3.2 Visual Question Answering

The task requires the model to predict an answer given an image and a question. We evaluate X^2-VLM on the VQA v2.0 dataset [65]. Following existing methods [7, 11, 5], we use both train and validation sets for training and include additional question-answer pairs from Visual Genome. Following ALBEF, we use a six-layer Transformer decoder to generate answers based on the outputs of the fusion module. Then, the model is fine-tuned by optimizing the auto-regressive loss. During inference, we constrain the decoder to only generate from the 3,129 candidate answers to make a fair comparison with existing methods. Note that there is a NULL answer. Thus, the actual number of

Method	# Params	VQA		NLVR2		RefCOCO+			COCO Caption	
		test-dev	test-std	dev	test-P	val	testA^d	testB^d	BLEU@4	CIDEr
Models Pretrained on COCO, VG, SBU and CC datasets (4M)										
ALBEF	210M	74.5	74.7	80.2	80.5	-	-	-	-	-
VLMo$_{\text{base}}$	175M	76.6	76.9	82.8	83.3	-	-	-	-	-
METER	341M	77.7	77.6	82.3	83.1	-	-	-	-	-
VL-BEiT	175M	77.5	77.8	81.9	82.7	-	-	-	-	-
X^2-VLM$_{\text{base}}$	255M	**79.2**	**79.3**	**85.9**	**86.1**	**85.4**	**89.2**	**77.3**	**41.0**	**133.6**
VLMo$_{\text{large}}$	562M	79.9	80.0	85.6	86.9	-	-	-	-	-
X^2-VLM$_{\text{large}}$	593M	**80.5**	**80.5**	**87.2**	**87.6**	**86.9**	**90.1**	**80.2**	**42.0**	**136.7**
Models Pretrained on More Data										
OmniVL	288M	78.3	78.4	-	-	-	-	-	39.8	133.9
SimVLM$_{\text{base}}$	273M	77.9	78.1	81.7	81.8	-	-	-	39.0	134.8
OFA$_{\text{base}}$	182M	78.0	78.1	-	-	81.4	87.2	74.3	41.0	**138.2**
BLIP$_{\text{base}}$	240M	78.2	78.2	82.5	83.1	-	-	-	39.4	131.4
X^2-VLM$_{\text{base}}$	255M	**80.4**	**80.2**	**86.2**	**87.0**	**85.2**	**90.3**	**78.4**	**41.7**	136.1
SimVLM$_{\text{large}}$	783M	79.3	79.6	84.1	84.8	-	-	-	40.3	**142.6**
OFA$_{\text{large}}$	472M	80.3	80.5	-	-	85.8	89.9	79.2	42.4	142.2
X^2-VLM$_{\text{large}}$	593M	**81.9**	**81.8**	**88.7**	**89.4**	**87.6**	**92.1**	**81.8**	**42.6**	139.1

Table 5: **Results on downstream image-text tasks**, including visual question answering (VQA), visual reasoning (NLVR2), visual grounding (RefCOCO+), and image caption generation (COCO Caption).

candidate answers is 3,128. Following previous work [22–24], the resolution of input images is set to 768x768.

We report the experimental results of VQA in Table 5. We can see that X^2-VLM$_{\text{base}}$ and X^2-VLM$_{\text{large}}$ outperforms other approaches with similar scale of model size. Specifically, X^2-VLM$_{\text{base}}$ substantially outperforms ALBEF, VLMo, METER, and VL-BEiT in the 4M setting. Besides, with more pre-training data, X^2-VLM$_{\text{base}}$ outperforms BLIP which also exploits large-scale image-text pairs from LAION. Compared to OmniVL which also supports both image-text tasks and video-text tasks, X^2-VLM$_{\text{base}}$ substantially outperforms it, achieving an absolute improvement of 2%. X^2-VLM also substantially outperforms SimVLM and OFA on both base and large scales. SimVLM utilizes an in-house 1.8B image-text dataset. OFA also leverages image annotations of objects and regions the same as X^2-VLM. These results confirm the effectiveness of the proposed framework for multi-grained vision language pre-training. Furthermore, when comparing X^2-VLM's performances in different settings in Table 5, we can see that the proposed framework has good scalability which can benefit from a larger model size. When pre-training a larger model with more data, the performance improvement is even more remarkable.

4.3.3 Visual Reasoning

We evaluate X^2-VLM on widely used benchmark NLVR2 [66]. The task lets the model determine whether a text describes the relations between two images. Following previous work [67, 35], we formulate the triplet input to two image-text pairs, each containing the text description and one image. We then concatenate the final output [CLS] features of the fusion module of the two pairs to predict the label. The resolution of input images is set to 384x384. Given the results in Table 5, we can observe that the visual reasoning task benefits more from the model size than the pre-training data scale. Comparing to other base-scale models, e.g. ALBEF, VLMo, VL-BEiT, SimVLM, and BLIP, X^2-VLM$_{\text{base}}$ has much better performance, achieving $\sim$ 3-4% absolute improvement, no matter when pre-training with 4M data or with much more noisy data. X^2-VLM$_{\text{large}}$ also substantially outperforms other large-scale models, including VLMo$_{\text{large}}$ and SimVLM$_{\text{large}}$.

	Winoground			OVAD			
	Group	Text	Image	All	Head	Medium	Tail
Random	12.5	25.0	25.0	8.6	36.0	7.3	0.6
CLIP	8.0	30.7	10.5	17.0	44.3	18.4	5.5
ALBEF_{4M}	11.0	29.2	15.5	15.6	43.1	17.3	3.7
BLIP	11.7	35.5	15.0	24.3	51.0	28.5	9.7
BLIP-2	18.2	43.0	22.0	-	-	-	-
$\text{UNITER}_{\text{large}}$	10.5	38.0	14.0	-	-	-	-
PEVL	12.2	33.2	15.7	-	-	-	-
OVADetector	-	-	-	21.4	48.0	26.9	5.2
$\textbf{X}^2\textbf{-VLM}_{\text{base4M}}$	22.5	46.3	25.3	24.0	51.9	29.7	7.2
$\textbf{X}^2\textbf{-VLM}_{\text{large4M}}$	25.5	49.5	31.0	27.7	54.0	34.4	10.1
$\textbf{X}^2\textbf{-VLM}_{\text{base}}$	24.5	47.3	29.8	27.6	52.2	34.7	10.3
$\textbf{X}^2\textbf{-VLM}_{\text{large}}$	**25.8**	**52.5**	**32.5**	**29.2**	**55.1**	**36.4**	**11.3**

Table 6: **Zero-shot evaluation results on fine-grained downstream tasks**: Winoground, a fine-grained image-text matching task, and OVAD, Open-vocabulary Attribute Detection(mAP).

4.3.4 Visual Grounding

We evaluate X^2-VLM on RefCOCO+ [54]. Given an image as the input and a text description as the query, the final output [CLS] features of the fusion module is utilized to predict the bounding box of the visual concept. The resolution of input images is set to 384x384. As indicated in Table 5, X^2-VLM outperforms OFA [22] which also utilizes image annotations of objects and regions for pre-training. Differently, OFA with an encoder-decoder architecture formulates all the data in the form of sequence-to-sequence. Furthermore, X^2-VLM for general V+L purposes outperforms MDETR [21] specialized for visual grounding tasks, achieving absolute improvements of $\sim 7\%$ (average on metrics). These results confirm the effectiveness of the proposed multi-grained vision language pre-training compared to other approaches that also leverage image annotations of objects and regions.

4.3.5 Image Captioning

The task requires a model to generate text descriptions of input images. Though X^2-VLM is more for cross-modal understanding, we also evaluate its generation performance on the COCO Captioning dataset [68]. Following UniLM [69] and BEiT-3, we use left-to-right MLM for generation. Specifically, we employ the text module and fusion module as decoder with left-to-right self-attention and adopt the method [70] that decreases finetune-generation discrepancy in MLM generation. The resolution of input images is set to 480x480. We report BLEU-4 and CIDEr scores on the Karparthy test split. As shown in Table 5, $\text{X}^2\text{-VLM}_{\text{base}}$ outperforms BLIP [20] and SimVLM [19] which are designed for generative tasks. BLIP exploits large-scale image-text pairs from LAION the same as X^2-VLM. SimVLM utilizes an in-house 1.8B image-text dataset. X^2-VLM also outperforms OFA [22] in image captioning in terms of BLEU-4. OFA has an encoder-decoder architecture and formulates all downstream tasks into sequence-to-sequence form for pre-training. In general, though X^2-VLM is more for cross-modal understanding, it performs competitively or sometimes better compared with SoTA generative methods.

4.3.6 Winoground

Winoground [71] presents a challenging task: given two images and two captions, the goal is to match them correctly, where the captions contain identical sets of words, but in a different order. Three metrics, namely Text (whether a model can match the correct caption for a given image), Image (vice versa), and Group (whether a model can match each pair), are used to evaluate the performance. Several competitive VLMs have been shown to perform close to or even below random chance. Experimental results in Table 6 shows that even when trained on 4M data X^2-VLM substantially outperforms other models such as $\text{UNITER}_{\text{large}}$, which is based on a large pre-trained object detector, and BLIP-2, which consists of giant ViT and FlanT5 large language model [72] and is pre-trained

Model	# Params	Video-QA		MSRVTT (1K test set)		
		MSRVTT	MSVD	R@1	R@5	R@10
ALPRO	513M	42.1	45.9	-	-	-
VIOLET	163M	43.9	47.9	-	-	-
All-in-one	110M	44.3	47.9	37.9	68.1	77.1
OmniVL	288M	44.1	51.0	47.8	74.2	83.8
$\mathbf{X^2}$**-VLM**$_{\text{base}}$	255M	45.0	52.8	47.6	74.1	84.2
$\mathbf{X^2}$**-VLM**$_{\text{large}}$	593M	**45.5**	**54.6**	**49.6**	**76.7**	**84.2**

Table 7: **Fine-tuning results on video-text tasks**, including video question answering on MSRVTT and MSVD datasets, and text-to-video retrieval on MSRVTT. We report classification accuracy for VQA and Recall@K for text-to-video retrieval.

on a much larger dataset with 129M images. Notably, the performance of X^2-VLM can be further improved by increasing the model size or pre-training dataset.

4.3.7 Open-vocabulary Attribute Detection

Open-Vocabulary Attribute Detection (OVAD) [73] aims to recognize an open set of objects in an image together with an open set of attributes for every object. We follow the benchmark and evaluate zero-shot performance of vision language models on attributes in the box-oracle setting. The experimental results are given in Table 6. X^2-VLM$_{\text{base}}$ pre-trained with 4M dataset has already been comparable to BLIP pre-trained with 129M dataset. X^2-VLM$_{\text{base}}$ also outperforms OVADetector which consists of a frozen CLIP text encoder and an object detector based on Faster-RCNN. Moreover, scaling X^2-VLM with larger pre-training datasets or larger model size consistently improve its performance as in other tasks.

4.4 Video-Text Downstream Tasks

X^2-VLM unifies image-text and video-text pretraining. In this section, we evaluate X^2-VLM on three widely used video-text tasks, including both **Video-Text Retrieval** (MSRVTT [74]) and **Video Question Answering** (MSRVTT-QA [75] and MSVD-QA [75]). We implement a text-to-video retrieval model the same as image-text retrieval by first calculating top-k candidates and then re-ranking the candidates using the fusion module. k is set to 32. During training and inference, we sample five frames for each video. The image resolution is set to 384. Video question answering requires a model to generate an answer given a video and a question. Following previous work, we formulate it as a classification task given candidate answers. During training and inference, we sample five frames for each video in the MSRVTT dataset, and eight frames for the MSVD dataset. The image resolution is set to 320 for MSRVTT and 224 for MSVD. We compare with SoTA video-language foundation models: ALPRO [41], VIOLET [42], and All-in-one [43]. We also compare X^2-VLM with OmniVL which also supports both image-text tasks and video-text tasks. There are other methods optimized specifically for either video-text retrieval [44, 45] or video question answering [46], which are not included in our comparison.

The results are given in Table 7. We can see that X^2-VLM$_{\text{base}}$ outperforms previous video-language foundation models on both video question answering and text-to-video retrieval, and X^2-VLM$_{\text{large}}$ further advance the performance, achieving new SoTA results of video-text pre-training. Besides, we compare X^2-VLM with OmniVL on both image-text (Table 3 and Table 5) and video-text benchmarks. In general, X^2-VLM$_{\text{base}}$ substantially outperforms OmniVL on all image-text downstream tasks, including image-text retrieval, visual question answering, image caption generation, and video question answering.

4.5 Multilingual Multi-modal Tasks

In X^2-VLM architecture, text encoding, vision encoding and fusion are separated. Accordingly, the capabilities of vision encoding and fusion would be kept when replacing the text encoder, leading to an efficient adaptation of the new text encoder. Our study demonstrates that we can replace the text encoder after pre-training on English data with a language-specific or domain-specific one to

Model	Flickr30K				MSCOCO		
	EN	DE	FR	CS	EN	ZH	JA
Models with Multilingual Multimodal Pretraining							
M^3P	87.7	82.7	73.9	72.2	88.7	86.2	87.9
UC^2	88.2	84.5	83.9	81.2	88.1	89.8	87.5
$MURAL_{base}$†	92.2	88.6	87.6	84.2	88.6	-	88.4
$MURAL_{large}$†	93.8	90.4	89.9	87.1	92.3	-	91.6
CCLM	96.0	93.3	93.7	92.8	94.1	93.0	94.3
X^2-VLM$_{base}$	96.7	94.0	93.5	92.9	94.9	93.0	95.2
X^2-VLM$_{large}$	**97.1**	**94.5**	**95.1**	**94.9**	**95.3**	**93.3**	**95.6**

Table 8: **Results on multilingual multi-modal tasks**. All the methods except X^2-VLM rely on data that are costly to collect to perform multilingual multi-modal pre-training. We evaluate model performance in English (EN), German (DE), French (FR), Czech (CS), Chinese (ZH), and Japanese (JA). Following previous work, we report the average Recall@K for both image-to-text retrieval and text-to-image retrieval with K = 1, 5, 10.

	Flickr30K		VQA	RefCOCO+		OVAD
	TR	IR	test-dev	testA^d	testB^d	mAP
Ours	**98.0**	89.0	**78.4**	**88.6**	**76.7**	**27.9**
w/o X^2-VLM	96.0	85.9	77.6	78.6	59.0	20.6
w/o multi-grained align	96.6	86.2	77.7	87.3	75.3	23.1
w/o bbox loss	97.4	**89.6**	78.2	83.6	66.0	26.8
w/o object data	97.2	86.8	78.1	88.1	76.5	26.4
w/o region data	97.8	89.0	78.0	84.8	69.3	22.3

Table 9: **Ablation study** of different components in the proposed framework and different types of data utilized.

support more applications in different languages or domains. Such a feature is hard to achieve with unified models like OFA and BEiT-3. For instance, BEiT-3 shares image, text, and fusion in a single Transformer, and thus replacing the text encoder can cause the capabilities of image encoding and fusion to be lost as well.

In this section, we replace the English text encoder of X^2-VLM with a multilingual text encoder XLM-R [25]. Then, without a second step multilingual multi-modal pre-training, we simply finetune X^2-VLM on multilingual multi-modal downstream tasks. We choose Multi30K [76] and multilingual MSCOCO [68, 77, 78] for evaluation since other multilingual multi-modal benchmarks such as IGLUE [79] do not have a training set. Following previous work, we compute the average Recall@K for both image-to-text retrieval and text-to-image retrieval with K = 1, 5, 10, as the evaluation metric.

We compare X^2-VLM with SoTA multilingual multi-modal pre-training methods. M^3P [49] utilizes 101G texts covering 100 languages. UC^2 [26] translates image-text pairs in English into five different languages. MURAL [27] collects large-scale image-text pairs in 110 languages. CCLM [28] utilizes parallel multilingual text pairs. All these methods rely on data that are costly to collect, while X^2-VLM relieves the multilingual multi-modal pre-training process. As shown in Table 8, X^2-VLM surprisingly outperforms all these methods in all six languages. The results indicate the potential of X^2-VLM being applicable to other domains or languages using a different text encoder without further pre-training.

4.6 Ablation Study

We conduct an in-depth ablation study and the results are given in Table 9. We describe the experimental settings in Appendix. First, we investigate the role of different components in the proposed framework and conduct an ablation of multi-grained aligning and box prediction loss

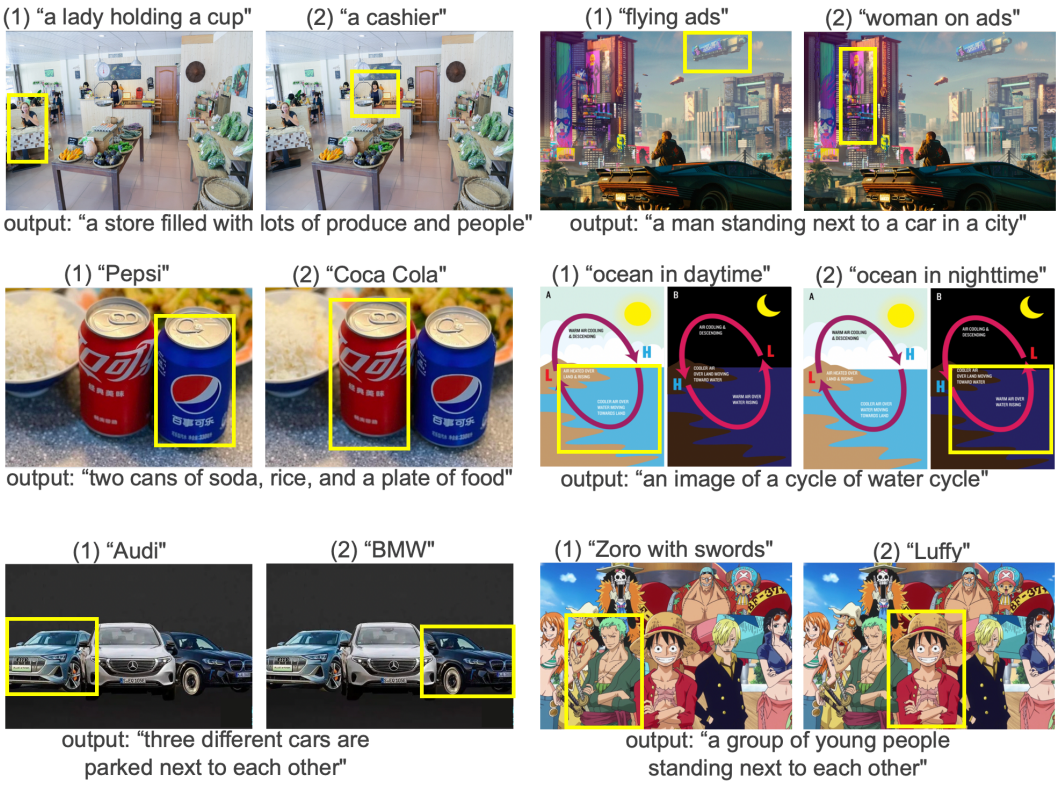


Figure 4: **Visualization of X^2-VLM generating image captions and locating visual concepts given manual input texts.** Only the image in the upper left corner is from the COCO dataset. Others are out-of-domain images from the internet. We give more examples in Appendix where we test X^2-VLM on images from robot grasping, e-commerce websites, and children's textbooks.

respectively. It should be noted that both object and region data are utilized in these two variants. The experimental results demonstrate that multi-grained aligning is more important for the model performance than the box prediction loss in all tasks, except the visual grounding task. The box prediction loss is critical to performance on visual grounding tasks, and combining the box prediction with multi-grained aligning further improves the model performances (Ours vs. w/o bbox loss).

Second, we explore the impact of different types of annotation data used in X^2-VLM, and ablate object data and region data respectively. Both multi-grained aligning and box prediction loss are applied in these two variants. The results indicate that both types of annotations are important to performance. Object data improve image-text retrieval, while region data are critical to visual grounding and open-vocabulary attribute detection. Combining object and region data yields the best performances (Ours vs. w/o object and w/o region). The w/o X^2-VLM variant, which ablates both multi-grained aligning and box prediction loss, or both object and region data, has the worst performances in all the tasks. We also provide an ablation study on temporal modeling methods in Appendix.

4.7 Qualitative Study of Multi-Grained Alignments

In this section, we provide a qualitative study of what vision language alignments have been learned by X^2-VLM. To this end, we ask X^2-VLM to generate image captions to see if it can describe an image appropriately. We also ask X^2-VLM to locate visual concepts in an image given manual input descriptions to see whether it can understand fine-grained objects or regions in an image. We use X^2-VLM$_{\text{large}}$ fine-tuned on COCO Caption and RefCOCO+ dataset respectively for this evaluation. We visualize the results in Figure 4, in which we choose some out-of-domain images from scientific posters, video games, cartoons, etc.

The visualization examples show that X^2-VLM can describe all these images appropriately with a precise understanding of the main characters or objects and their relationships. When asking X^2-VLM to locate visual concepts in an image according to the descriptions we provided, we find that it can capture small objects in the background or objects which have been partially obscured. Moreover, X^2-VLM can recognize different brands of soda or cars or distinguish "Luffy" and "Zoro" from other cartoon characters. We give more examples in Appendix, where X^2-VLM can also recognize "Albert Einstein", "Edison", "Ultraman", and "Doraemon". It is surprising since the annotations of objects or regions we exploited in pre-training are only about common objects such as "soda", "car", or "man". The results indicate that X^2-VLM learns to localize diverse fine-grained visual concepts from large-scale noisy image-text pairs.

5 Conclusion and Discussion

In this paper, we have proposed to learn multi-grained alignments between vision and language in pre-training. To this end, we have proposed a unified framework for multi-grained vision language pre-training that directly aligns the multi-grained vision features with the paired text features and simultaneously locates multi-grained visual concepts in the same image given different text descriptions. Based on it, we have presented X^2-VLM, an all-in-one pre-trained VLM with a flexible modular architecture, in which we have further unified image encoding and video encoding to make it able to handle both image-text tasks and video-text tasks.

We have conducted extensive experiments to verify the effectiveness of X^2-VLM. The results have shown that X^2-VLM substantially outperforms SoTA image-text pre-training methods on base and large scale in many downstream image-text tasks, making a good trade-off between performance and model scale. X^2-VLM is also the new SoTA pre-trained model on video-text tasks, including video-text retrieval and video VQA. Experimental results also show that the proposed framework for multi-grained vision language pre-training is scalable to massive data and a larger model size. Moreover, we have revealed the potential of the modular design of X^2-VLM, showing it can be utilized in other languages or domains. By replacing the text encoder with XLM-R after pre-training on English data, X^2-VLM outperforms SoTA methods on multi-lingual multi-modal tasks.

We also have provided an in-depth ablation study to investigate the role of different components in the proposed framework. Experimental results have shown that both multi-grained localization and multi-grained aligning are critical components of the proposed method. Furthermore, we have conducted a qualitative study of what vision language alignments have been learned by X^2-VLM. We have found that by training with large-scale image-text pairs, X^2-VLM learns to locate diverse fine-grained visual concepts in open-domain images, such as different brands of sodas, cars, and characters or celebrities.

Acknowledgements

We sincerely thank our colleagues at ByteDance, Tao Kong, for his constructive and detailed feedback on this work, and Jiaze Chen for his generous assistance in the training of X^2-VLM.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[3] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

[4] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[5] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 2021.

[6] Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al. Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*, 2021.

[7] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514.

[8] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 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 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23, 2019.

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

[10] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[11] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[12] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[13] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[15] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8429–8438. IEEE, 2019. doi: 10.1109/ICCV.2019.00852.

[16] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018.

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[19] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.

[21] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

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

[23] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[24] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

[25] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020.

[26] Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4155–4165, 2021.

[27] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. Mural: multimodal, multitask retrieval across languages. *ArXiv preprint*, abs/2109.05125, 2021.

[28] Yan Zeng, Wangchunshu Zhou, Ao Luo, and Xinsong Zhang. Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training. *arXiv preprint arXiv:2206.00621*, 2022.

[29] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99, 2015.

[30] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00636.

[31] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10267–10276, 2020.

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[33] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022.

[34] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Zicheng Liu, Michael Zeng, et al. An empirical study of training end-to-end vision-and-language transformers. *arXiv preprint arXiv:2111.02387*, 2021.

[35] Hangbo Bao, Wenhui Wang, Li Dong, and Furu Wei. Vl-beit: Generative vision-language pretraining. *arXiv preprint arXiv:2206.01127*, 2022.

[36] Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. E2E-VLP: End-to-end vision-language pre-training enhanced by visual learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 503–513, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.42.

[37] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[38] Yongfei Liu, Chenfei Wu, Shao-yen Tseng, Vasudev Lal, Xuming He, and Nan Duan. Kd-vlp: Improving end-to-end vision-and-language pretraining with object knowledge distillation. *arXiv preprint arXiv:2109.10504*, 2021.

[39] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021.

[40] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021.

[41] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4953–4963, 2022.

[42] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021.

[43] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*, 2022.

[44] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022.

[45] Shaobo Min, Weijie Kong, Rong-Cheng Tu, Dihong Gong, Chengfei Cai, Wenzhe Zhao, Chenyang Liu, Sixiao Zheng, Hongfa Wang, Zhifeng Li, et al. Hunyuan_tvr for text-video retrivial. *arXiv preprint arXiv:2204.03382*, 2022.

[46] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *arXiv preprint arXiv:2206.08155*, 2022.

[47] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. *arXiv preprint arXiv:2209.07526*, 2022.

[48] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[49] Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3977–3986, 2021.

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[51] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 658–666. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00075.

[52] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 1143–1151, 2011.

[53] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238.

[54] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.

[55] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[56] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.303.

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

[58] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[59] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019.

[60] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021.

[61] Tiannan Wang, Wangchunshu Zhou, Yan Zeng, and Xinsong Zhang. Efficientvlm: Fast and accurate vision-language models via knowledge distillation and modal-adaptive pruning. *arXiv preprint arXiv:2210.07795*, 2022.

[62] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

[63] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298932.

[64] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

[65] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.670.

[66] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 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, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1644.

[67] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021.

[68] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[69] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems*, 32, 2019.

[70] Yan Zeng and Jian-Yun Nie. An investigation of suitability of pre-trained language models for dialogue generation – avoiding discrepancies. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4481–4494, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.393.

[71] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

[72] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[73] Maria A Bravo, Sudhanshu Mittal, Simon Ging, and Thomas Brox. Open-vocabulary attribute detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7041–7050, 2023.

[74] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[75] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.

[76] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166.

[77] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2066.

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

[79] Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. Iglue: A benchmark for transfer learning across modalities, tasks, and languages. *ArXiv preprint*, abs/2201.11732, 2022.

A Appendix

A.1 Pre-training Datasets

As follows, we give some data filtering details. Since LAION and the video-text datasets are too large, we have filtered the datasets to speed up the pre-training. Specifically, for LAION, we use English data only. Following BLIP [20], we remove an image if the shorter edge is smaller than 224 pixels. We also remove an image if the ratio of height/width or width/height is larger than 3. For video clip-text pairs, we remove a pair if the number of words is less than 2. Following previous work, we use CLIP score to filter video data. We sample a frame for a video clip and we calculate the CLIP score between the frame and the text. We remove a video clip-text pair if the score is less than 0.25. For image annotations of objects and regions, we remove a sample because of: 1) invalid annotations (e.g. negative values for bounding boxes or boxes being outside of the images); 2) boxes being too small (less than a patch); 3) highly overlapped text descriptions of regions ($> 75\%$), etc. For an object annotation, if it contains an object attribute, e.g. color, we concatenate the attribute with the object label as the text description. Moreover, some images in the OpenImages dataset contain relationship annotations, indicating pairs of objects in particular relations (e.g. "woman playing guitar", "beer on table"), object properties (e.g. "table is wooden"), and human actions (e.g. "woman is jumping"). We also utilize this part of data.

A.2 Implementation Details

X^2-VLM is pre-trained at image resolution of 224×224 using 16×16 patch size. Though, as indicated in previous work such as OFA [22] and CoCa [23], increasing resolution will improve model performance, we keep it small to accelerate pre-training. Besides, we apply mixed precision for training. For text input, we set the maximum number of tokens to 30. To further speed up pre-training with large-scale data, we divide the training process into two steps. First, we train X^2-VLM with large-scale image-text pairs. Then, we further train X^2-VLM on video-text pairs and the 4M dataset. The reason behind this is that training on video data is slow. Because of it, we randomly sample only three frames for a video clip in pre-training. We mix all types of data in a training batch, and thus for each training iteration, we optimize the model by multi-grained aligning and multi-grained localization simultaneously.

With 4M data, we pre-train X^2-VLM$_{\text{base}}$ for 500K steps with a batch size of 1024 on 8 A100 and X^2-VLM$_{\text{large}}$ for 250K steps on 16 A100, which takes ~ 1 week. The learning rate of X^2-VLM$_{\text{base}}$ is warmed-up to $1e^{-4}$ in the first 2500 steps and decayed following a linear schedule. The learning rate is $5e^{-5}$ for X^2-VLM$_{\text{large}}$. With large-scale data, training X^2-VLM takes 2-3 weeks on 32 A100 for the base model and 64 A100 for the large model.

A.3 Ablation Study

To ensure a fair comparison, all compared model variants are trained on 4M images for 100K steps. Following previous studies, we have shortened the training steps to compare different ablated variants more efficiently. We evaluate model performance on image-text retrieval (Recall@1), visual question answering, visual grounding, and zero-shot open-vocabulary attribute detection. It is worth noting that VQA has a large train and test set, which means that even a relatively small difference in performance is worth considering.

	Flickr30K		**VQA**	**MSRVTT**	**Video-QA**	
	TR	IR	test-dev	IR	MSRVTT	MSVD
w/ avg pool (ours)	**98.5**	**90.4**	**80.4**	**47.6**	**45.0**	**52.8**
w/ temporal attn	98.2	89.6	80.0	45.6	44.4	52.1

Table 10: **Ablation study** of temporal modeling methods.

Additionally, we investigate whether better temporal modeling could further improve video understanding capabilities while maintaining good image understanding, as presented in Table 10. We use an established method that adds temporal attention in ViT. The experimental results on image/video-text retrieval and image/video VQA show that simply averaging the features of each frame achieves

better performances on all tasks. We suppose that our approach is more unified in modeling both image and video features, and thus strong image understanding capability is better transferred to video understanding.

A.4 Qualitative Study of Multi-Grained Alignments

Figure 5: Visualization of X^2-VLM generating captions for images and locating visual concepts given manual input descriptions.

We provide a qualitative study of what vision language alignments have been learned by X^2-VLM. To this end, we ask X^2-VLM to generate image captions or to locate visual concepts. We visualize the results in Figure 5, where the first two images are from the in-domain COCO dataset. We find that X^2-VLM can capture small objects in the background or objects which have been partly masked. We also choose out-of-domain images for evaluation. As shown in Figure 6, Figure 7, Figure 8, and Figure 9, X^2-VLM can recognize many visual concepts from different domains.

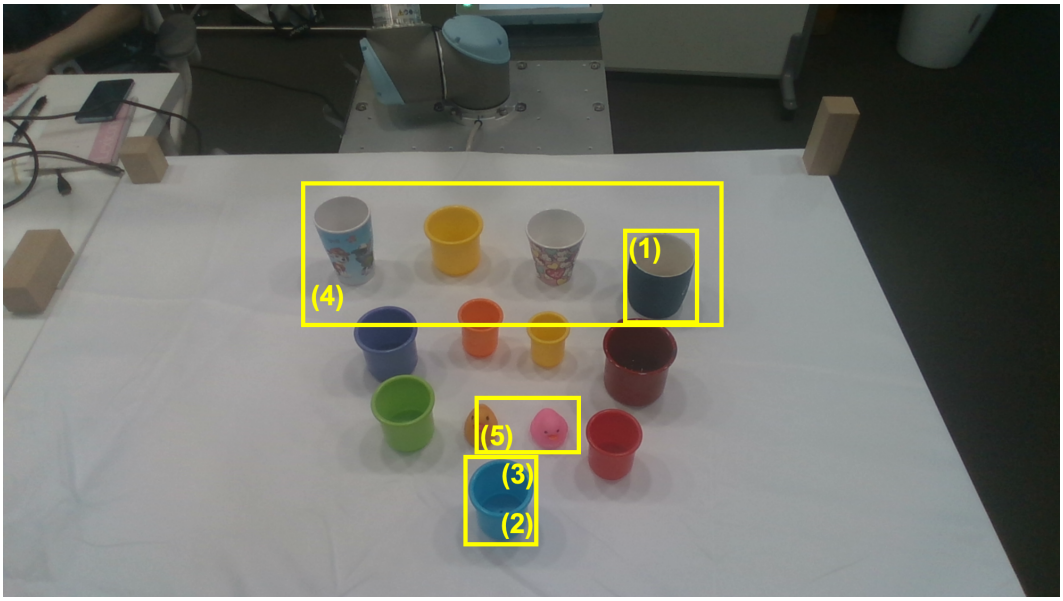


Figure 6: Visualization of X^2-VLM locating visual concepts in robot grasping scene given text descriptions: 1) "deep blue cup"; 2) "light blue cup"; 3) "blue cup at the bottom"; 4) "four cups at the top"; 5) "two small ducks".

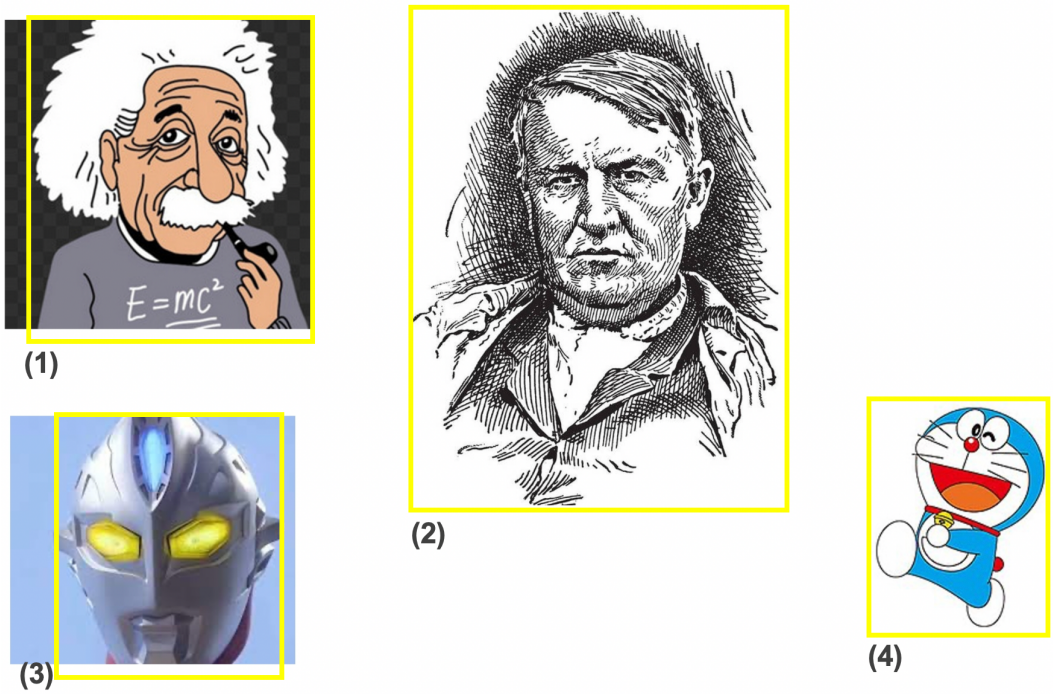


Figure 7: Visualization of X^2-VLM locating celebrities and cartoon characters given text descriptions: 1) "Albert Einstein"; 2)"Edison"; 3)"Ultraman"; 4)"Doraemon".

Figure 8: Visualization of X^2-VLM locating objects in an image from an e-commerce website in China. The input text descriptions are: 1) "shoes"; 2)"vacuum cleaner"; 3)"lipstick"; 4)"dress".

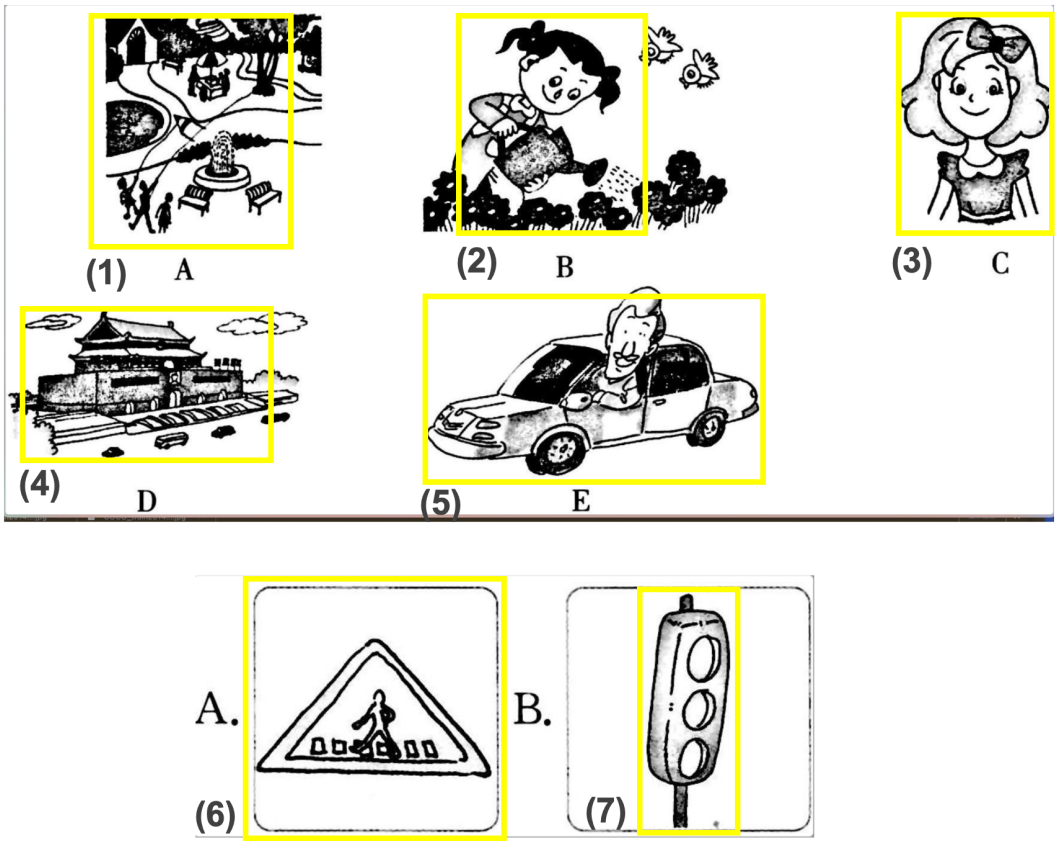


Figure 9: Visualization of X^2-VLM locating visual concepts in an image from children's textbooks. The input text descriptions are: 1) "flying kites in the park"; 2) "watering flowers"; 3) "well dressed girl"; 4) "Tiananmen Tower"; 5) "drive to work"; 6) "sign"; 7) "traffic lights".