CommerceMM: Large-Scale Commerce MultiModal Representation Learning with Omni Retrieval

Licheng Yu
lichengyu@fb.com
Meta AI
Menlo Park, CA, USA

Jun Chen
junchen20@fb.com
Meta AI
Menlo Park, CA, USA

Animesh Sinha
animeshsinha@fb.com
Meta AI
Menlo Park, CA, USA

Mengjiao MJ Wang
mengjiaow@fb.com
Meta AI
Menlo Park, CA, USA

Yu Chen
Meta AI
Menlo Park, CA, USA
hugochen@fb.com

Tamara L. Berg
tlberg@fb.com
Meta AI
Menlo Park, CA, USA

Ning Zhang
ningzhang@fb.com
Meta AI
Menlo Park, CA, USA

ABSTRACT

We introduce CommerceMM - a multimodal model capable of providing a diverse and granular understanding of commerce topics associated to the given piece of content (image, text, image+text), and having the capability to generalize to a wide range of tasks, including Multimodal Categorization, Image-Text Retrieval, Query-to-Product Retrieval, Image-to-Product Retrieval, etc. We follow the pre-training + fine-tuning training regime and present 5 effective pre-training tasks on image-text pairs. To embrace more common and diverse commerce data with text-to-multimodal, image-to-multimodal, and multimodal-to-multimodal mapping, we propose another 9 novel cross-modal and cross-pair retrieval tasks, called Omni-Retrieval pre-training. We also propose a novel approach of modality randomization to dynamically adjust our model under different efficiency constraints. The pre-training is conducted in an efficient manner with only two forward/backward updates for the combined 14 tasks. Extensive experiments and analysis show the effectiveness of each task. When combining all pre-training tasks, our model achieves state-of-the-art performance on 7 commerce-related downstream tasks after fine-tuning.

CCS CONCEPTS

• **Computing methodologies** → *Neural networks*; • **Information systems** → *Multimedia and multimodal retrieval; Online shopping.*

KEYWORDS

multimodal, vision and language, representation learning

1 INTRODUCTION

At Meta, nearly every post related to commerce is multimodal, e.g., a Marketplace[1] post is composed of one or several views of a product associated with its product description, a Shop product listing is composed of the product images and detailed specifics describing the product, e.g., title, attribute, size, material, and influencers upload their fashion posts to Instagram[2] with captions and hashtags. Since the announcement of Facebook/Instagram Shops[3] on Meta Business, there has been an exploding increase of usage of our commerce platforms. For example, users perform text query search on Marketplace every second looking for specific products, users also tag the products (e.g., t-shirt, necklace) while uploading a post, they also trigger visual search looking for similar products appearing in other images on the platform, etc. The large-scale commerce-related data and various use cases motivate us to build a commerce-specific multimodal representation for the post.

Recently vision-and-language representation learning is becoming a more and more popular research topic. This trend has also motivated people to study the commerce-specific pre-training [8, 11, 49, 50]. In those works, the authors pre-train the transformer-based [37] model on commerce image-text pairs, then fine-tune it on image-text retrieval, image captioning, category recognition, etc. However, most of existing works were trained on the medium-scale image-text pairs and evaluated on limited academic tasks. This raises a question: can Meta learn a more generalized multimodal representation for various practical commerce-related applications?

[1] https://www.facebook.com/marketplace
[2] https://www.instagram.com
[3] https://www.facebook.com/business/shops

Figure 1: Example cross-modal and cross-pair data. Users could use text query to do product search. Some users tag the relevant products when uploading their multimodal media. On product page, there could be multiple views of products. While those medias are of different type (text, multimodal, image), they are linked with the same product.

In this spirit, we introduce **Commerce MultiModal** Representation (**CommerceMM**), a large-scale pre-trained model for joint multimodal commerce embedding at Meta. We scale up the pre-training data from less than 1 million to hundreds of million. Our model is composed of an image encoder, a text encoder (transformer-based), and a multimodal fusion encoder (transformer-based), as in Fig. 2. We then propose two sets of pre-training tasks for the model and evaluate on a wide range of 7 commerce-related tasks.

The first set of pre-training tasks consists of 5 effective ones on the image-text pairs, including Masked Language Modeling (MLM), Masked Image Modeling KL-Divergence (MIM-kl), Masked Image Modeling Feature Regression (MIM-fr), Image-Text Contrastive Learning (ITC), and Image-Text Matching (ITM). While previous works [10, 16] show MIM is not helpful for multimodal pre-training and typically MLM is the most effective task [5, 19], our proposed MIM tasks are essential. The key difference lies in (*i*) we use a larger mask ratio of 50% instead of 15% used in [5, 25, 36]; (*ii*) we only recover the signal at image's [CLS] token instead of reconstructing every masked patch; (*iii*) the supervision comes from the other two views (ClusterFit [43] and GrokNet [4]) of the intact raw image.

To our knowledge, all existing multimodal pre-training take the image-text pairs as the default input for the representation learning. However, in practice, specifically in the commerce domain, we have access to more diverse formats of data, *e.g.*, multiple images, texts, and other valuable data, e.g., query and click, around the same product. For example, in Fig. 1(a), we show users could perform query search to find the related product. User's click on some product builds a text-to-multimodal mapping between the query text and the multimodal product page. Fig. 1(b) shows an example of users tagging the product when uploading their own multimodal media consisting of a photo and a caption, which builds a multimodal-to-multimodal mapping. Fig. 1(c) shows a product page with multiple views of the same product, where each of them could be a source image to search for the associated product, making an image-to-multimodal mapping. We call such data as cross-modal and cross-pair data. Apparently, the common image-text pre-training tasks like ITC and ITM can no longer be applied since their input is fixed to be the image-text pair and can only handle the image-text relation. This motivates us to propose the second set of pre-training tasks - Omni-Retrieval Pre-training. We propose to build the relations between any modality type from image, text, and multimodality. Specifically, we encapsulate the cross-pair data into the form of (source image, source text, target image, target text). We predict three embeddings for the source and target respectively: visual embedding, textual embedding, and multimodal embedding. Exhaustive matching results in 9 types of pairwise relations, where each relation can be learned with contrastive learning mimicking 9 retrieval tasks in total. With the 9 Omni-Retrieval tasks, our pre-trained model is shown to learn a more discriminative and generalized representation for any alignment space.

We combine both 5 image-text tasks and 9 omni-retrieval tasks together during pre-training. While the combined pre-training (in total 14 tasks) seem unbelievably heavy at first glance, our pre-training is performed in a simple and efficient manner. During training, there is only 1 forward/backward for each set of tasks, and we found the two sets of tasks are complementary with each other.

Besides pre-training, we propose another novel idea – modality randomization. We change the text encoder and multimodal fusion encoder layers dynamically during training. At each training step, we randomly assign K and M transformer layers (while keeping the sum of K and M unchanged) to the text and multimodal encoder. The shuffling role makes each layer learn from both modalities and allows knowledge sharing. Our experiments show a light-weighted 2nd-stage pre-training with modality randomization further improves multimodal understanding. Additionally, our model can be flexibly adjusted to different architectures per different efficiency constraints.

We introduce 7 diverse commerce tasks to evaluate the pre-trained model, including Catalog Categorization, Marketplace Categorization, Image-to-Text Retrieval, Text-to-Image Retrieval, Query-to-Product Retrieval, Image-to-Product Retrieval, and Image-to-Image Retrieval. We present extensive experiments and analyses to provide useful insights on the effectiveness of each pre-training task and dynamic model architecture. The experiments show that with all 14 pre-training tasks, our model can achieve state-of-the-art performance across all 7 downstream tasks. Furthermore, with dynamic model randomization, we can achieve an even better performance with a smaller model size. Our contributions are summarized as follows: (*i*) We propose 5 effective image-text pre-training tasks on a large-scale commerce multimodal dataset. (*ii*) We introduce a novel pre-training approach of Omni-Retrieval for the cross-modal and cross-pair data. (*iii*) We validate our approach on 7 diverse tasks with thorough study and insightful analysis.

2 RELATED WORK

Inspired by the success of BERT [7] in natural language processing, there has been a surge of interest in vision-and-language representation learning, by pre-training on the large-scale multimodal data with a transformer-based [37] model architecture, then fine-tuning on various downstream tasks [5, 18–22, 25, 34–36, 38–40, 47, 48]. While different works use different architectures (*e.g.* two-stream [25, 36, 45] *vs.* single-stream [5, 18, 21, 34]), features (*e.g.* regions [2] *vs.* grids [14]), backbones (*e.g.* ConvNets [14] *vs.* Transformers [16]) *etc.*, the shared goal of visual-linguistic pre-training is to exploit large-scale, paired image-and-text corpora [15, 17, 24, 28, 30, 32] and obtain models that are *implicitly* pre-built with the appropriate correspondences to understand the multi-modal data. The level of understanding is benchmarked with multiple multimodal downstream tasks [3, 29, 46].

Various tasks have been introduced for the multi-modal pre-training. The most notable three are Masked Language Modeling (MLM), Masked Image Modeling (MIM), and Image-Text Matching (ITM) – which are the direct counterparts of the BERT objectives. Several other variants of these three tasks have also been explored, such as predicting object tags [13, 23], masked region classification [5, 25], sequential caption generation [40, 48], image-text contrastive learning [15, 19, 30], etc. Recent works [10, 16] found MIM may not be essential for pre-training among those tasks. In this work, we found that our proposed MIM is more effective than MLM. Inspired by recent works of self-supervised learning on vision [12, 42], we propose to mask out image patches with larger proportion and follow MaskFeat [41] to reconstruct other views of the whole image rather than recovering those masked regions only.

Besides, all existing pre-training works are conducted on the image-text pairs. We can certainly convert the cross-modal and cross-pair data in Fig. 1 to the image-text pairs, but that would constraint the model from learning more general representations in other cross-modal space, e.g., multimodal-to-multimodal alignment. In comparison, we propose a novel approach of Omni-Retrieval pre-training to build the relations between any modality towards learning a universal representation.

3 MODEL AND PRE-TRAINING

3.1 Model Overview

The model architecture of CommerceMM is illustrated in Fig. 2(a), which is composed of an image encoder, text encoder, and a multimodal fusion encoder. Given an image, the vision encoder converts the raw pixels into a sequence of visual embeddings, *i.e.*, flattened grid/patch-wise features. On the other side, the text encoder converts the input sentence into a sequence of word tokens and feeds them into the transformer to get textual embeddings. Inside both visual and textural embeddings, there is a special [CLS] token, which encodes the whole image and sentence representation. We feed both sequences of embeddings into another transformer - multimodal fusion encoder, where the two modalities learn to interact with each other building a contextualized multimodal embeddings. We still keep the [CLS] tokens of both modalities at the output of multimodal fusion encoder.

Our image encoder is an off-the-shelf ResNet or ViT-B/16 model, that have been weakly-supervised trained on billions of Instagram

images using hashtags [26]. Our text encoder and multimodal fusion encoder are initialized from XLM-R [6], an unsupervised multilingual language model. Specifically, the text encoder inherits its first K layers and the multimodal fusion encoder inherits its remaining M layers, whose sum equals to XLM-R's total layers (6, 12, or 24 depending on the chosen model size). We keep K and M unchanged under most setting for a fixed model architecture.

Interestingly, such design also allows us to dynamically change K and M during training, *i.e.*, each layer shuffles between text encoder and multimodal encoder. In Sec. 4.6, we show the modality randomization can further enhance the multimodal understanding and make the model adjustable per different efficiency constraints.

Given an image-text pair, we denote its raw visual inputs as $\mathbf{v}_i = \{\mathbf{v}_i^{\text{cls}}, \mathbf{v}_i^1, ..., \mathbf{v}_i^K\}$, and its input words as $\mathbf{w}_i = \{\mathbf{w}_i^{\text{cls}}, \mathbf{w}_i^1, ..., \mathbf{w}_i^T\}$, where the subscript i indicates the i-th pair in the dataset. As above, an additional special [CLS] token is inserted to each sequence.

We introduce two sets of pre-training tasks: (*i*) Image-Text Pre-training, which consists of 5 tasks of MLM, MIM-kl, MIM-fr, ITC, and ITM as Fig. 2(a); (*ii*) Omni-Retrieval Pre-training, which consists of 9 cross-modal and cross-pair retrieval tasks as Fig. 2(b). We denote $\mathbf{m}$ as the mask indices, which is applied to mask out part of the sequences for self-supervised learning.

3.2 Image-Text Pre-training

Masked Language Modeling (MLM) In MLM, we randomly mask out the input words with probability of 15% and replace the masked ones $\mathbf{w}^{\mathbf{m}}$ with special token [MASK]. The goal of MLM is to predict these masked words based on their surrounding words $\mathbf{w}^{\backslash \mathbf{m}}$ and the visual context $\mathbf{v}$, by minimizing the negative log-likelihood:

$$\mathcal{L}_{\text{MLM}}(\theta) = -\mathbb{E}_{(\mathbf{w},\mathbf{v}) \sim D} \log P_\theta(\mathbf{w}^{\mathbf{m}} | \mathbf{w}^{\backslash \mathbf{m}}, \mathbf{v}), \quad (1)$$

where θ is the model parameters. Each pair $(\mathbf{w}_i, \mathbf{v}_i)$ is sampled from the whole training set D.

Masked Image Modeling (MIM) Similar to MLM, we also sample and mask the visual inputs, e.g., patches. Previous works [10, 16] found the naive masked patch regression is not helpful in multi-modal pre-training. In this work, we disregard the reconstruction of each masked region, but instead recover the holistic image signal at $\mathbf{v}^{\text{cls}}$ token. We first follow [12, 42] to use a larger masking ratio of 50% (instead of 15% as in [5, 25, 36]). The masked patches are replaced with grey pixels. Our supervision is provided by another view of the original intact input image. While MaskFeat [41] used Histograms of Oriented Gradients (HOG) as the supervision for visual pre-training, we rely on the more discriminative signals from ClusterFit [43] and GrokNet [4] to extract additional two views of the raw image. Between the two, [43] provides the clustering probability $c(\mathbf{v})$ while [4] extracts the pool5 embedding $r(\mathbf{v})$ (feature output from the 5-th Conv Block). Correspondingly, we propose two MIM tasks but sharing the same objective base:

$$\mathcal{L}_{\text{MIM}}(\theta) = \mathbb{E}_{(\mathbf{w},\mathbf{v}) \sim D} d_\theta(\mathbf{v}^{\mathbf{m}} | \mathbf{v}^{\backslash \mathbf{m}}, \mathbf{w}), \quad (2)$$

where d_θ is defined as follows.

1) **Masked Image Modeling Feature Regression (MIM-fr)** MIM-fr learns to regress the multimodal transformer's output at $\mathbf{v}_i^{\text{cls}}$ to the pool5 embedding from GrokNet, *i.e.*, $r(\mathbf{v})$. Specifically, we apply an FC layer to convert its hidden output into a vector $\phi_\theta(\mathbf{v}_i^{\text{cls}})$ of the same dimension as GrokNet's pool5. Then we

(a) CommerceMM on the input (image, text) pair and 5 image-text pre-training tasks.

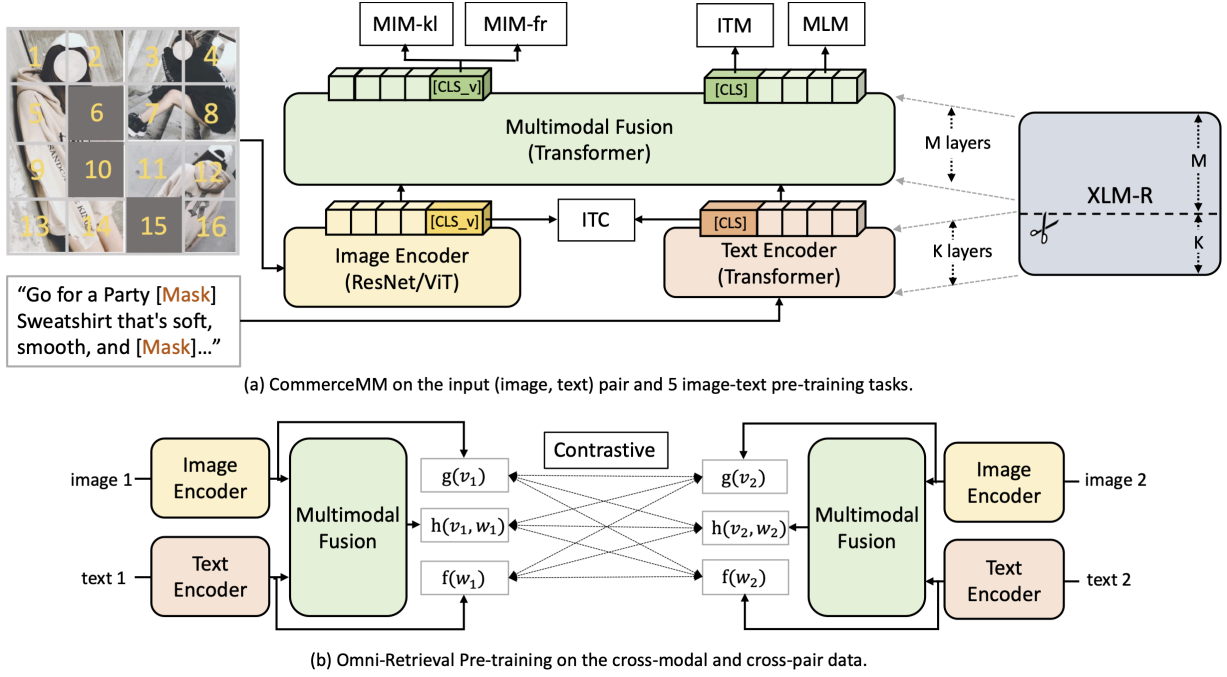(b) Omni-Retrieval Pre-training on the cross-modal and cross-pair data.

Figure 2: CommerceMM Model Architecture with the image-text pre-training and omni-retrieval tasks.

apply L2 regression on the mean square error between the two: $d_\theta(\mathbf{v}^{\text{cls}}|\mathbf{v}^{\backslash\mathbf{m}}, \mathbf{w}) = \|\phi_\theta(\mathbf{v}^{\text{cls}}) - r(\mathbf{v})\|_2^2$.

2) **Masked Image Modeling KL-Divergence (MIM-kl)** MIM-kl applies the soft label of ClusterFit probability[43] as the supervision signal, which is the softmax from ClusterFit output formatting the distribution of clusters $\tilde{c}(\mathbf{v})$. We project the multimodal transformer's output at $\mathbf{v}_i^{\text{cls}}$ to the same distribution space $\psi_\theta(\mathbf{v}_i^{\text{cls}})$ with softmax. We aim to distill the intact knowledge from ClusterFIT into CommerceMM, by minimizing the KL divergence between the two distributions: $d_\theta(\mathbf{v}^{\text{cls}}|\mathbf{v}^{\backslash\mathbf{m},\mathbf{w}}) = \|D_{KL}(\tilde{c}(\mathbf{v})\|\psi_\theta(\mathbf{v}^{\text{cls}}))\|$.

Image-Text Contrastive Learning (ITC) Following [19], we add an image-text contrastive loss between the visual and textual embeddings right before feeding them into the multimodal fusion module. It aims to align the two modalities into the same space before fusion. Specifically, we project the textual embedding $\mathbf{w}_i^{\text{cls}}$ and visual embedding at $\mathbf{v}_i^{\text{cls}}$ to a normalized lower-dimensional representations via two linear transformations f and g. The similarity of the text $\mathbf{w}_i$ and the image $\mathbf{v}_j$ is then measured by the dot product of $s(\mathbf{w}_i, \mathbf{v}_j) = f_\theta(\mathbf{w}_i^{\text{cls}})^T g_\theta(\mathbf{v}_i^{\text{cls}})$. We apply contrastive learning to bring the matched image-text pairs in the embedding space closer than the unmatched ones as follows:

$$\mathcal{L}_{itc} = -\sum_i [\log \frac{\exp(s(\mathbf{w}_i, \mathbf{v}_i)/\tau)}{\sum_j \exp(s(\mathbf{w}_i, \mathbf{v}_j)/\tau)} + \log \frac{\exp(s(\mathbf{w}_i, \mathbf{v}_i)/\tau)}{\sum_j \exp(s(\mathbf{w}_j, \mathbf{v}_i)/\tau)}],$$

where τ is a learned temperature parameter.

Image-Text Matching (ITM) In ITM, the inputs are a paired sentence and image and the output is a binary label $y \in \{0, 1\}$, indicating if each input pair is a match. We extract the hidden output of $\mathbf{w}^{\text{cls}}$ at the last layer of multimodal fusion to represent the fused representation of both modalities, then feed it into an

FC layer followed by a sigmoid function to predict a single score between 0 and 1. We denote the output score as $o_\theta(\mathbf{w}, \mathbf{v})$. During pre-training, we sample a positive or negative pair $(\mathbf{w}, \mathbf{v})$ from the dataset D at each step. The negative pair is created by replacing the image or text in a paired sample with a randomly-selected one from other samples. Following [19], when ITC is applied, we could sample the hard negative pairs from ITC's computed similarity matrix $s(\mathbf{w}_i, \mathbf{v}_j)$. The incorporation of those hard negatives makes ITM a harder task, which is more beneficial for the pre-training [27]. We apply binary cross entropy for this loss:

$$\mathcal{L}_{\text{ITM}}(\theta) = -\mathbb{E}_{(\mathbf{w},\mathbf{v})\sim D}[y \log o_\theta(\mathbf{w}, \mathbf{v}) + (1 - y) \log(1 - o_\theta(\mathbf{w}, \mathbf{v}))]).$$

3.3 Cross-Pair Pre-training: Omni Retrieval

As in Fig. 1, besides image-text pairs, there are also a huge amount of cross-modal and cross-pair commerce data. We formulate such data as two pairs, where we denote the source pair as $(\mathbf{w}_i, \mathbf{v}_i)$ and the target pair as $(\mathbf{w}_j, \mathbf{v}_j)$. Note one of $\mathbf{w}$ and $\mathbf{v}$ (in the source/target pair) could be missing in some case. For example, the source for search query is only one single sentence, and the source for visual search is only one single image, while both are linked with some multimodal product pages. We replace the missing image or text with grey pixels or an empty string and introduce an indicator δ to tell the existence of each modality. We replicate our model for both source and target pairs sharing the same parameters, as in Fig. 2(b). We first feed the source pair $(\mathbf{w}_i, \mathbf{v}_i)$ to our model, our image encoder, text encoder, and multimodal fusion return three embeddings at their corresponding [CLS] respectively. With 3 simple linear transformations, we get three normalized embeddings

$g(\mathbf{v}_i)$, $f(\mathbf{w}_i)$, and $h(\mathbf{w}_i, \mathbf{v}_i)$. Similarly, we can get the image embedding $g(\mathbf{v}_j)$, text embedding $f(\mathbf{w}_j)$, and multimodal embedding $h(\mathbf{w}_j, \mathbf{v}_j)$ for the target pair. If a source pair is linked with a target pair, we assume any existing modality from the source would be highly correlated with every existing modality from the target. Thus we compute the similarity score between any pair of source and target embedding from the text, image, or multimodal perspective respectively as the follows:

$$\text{Text:} \begin{cases} s_{ij}^{T \leftrightarrow T} = f(\mathbf{w}_i)^T f(\mathbf{w}_j) \\ s_{ij}^{T \leftrightarrow I} = f(\mathbf{w}_i)^T g(\mathbf{v}_j) \\ s_{ij}^{T \leftrightarrow M} = f(\mathbf{w}_i)^T h(\mathbf{w}_j, \mathbf{v}_j), \end{cases} \quad \text{Image:} \begin{cases} s_{ij}^{I \leftrightarrow T} = g(\mathbf{v}_i)^T f(\mathbf{w}_j) \\ s_{ij}^{I \leftrightarrow I} = g(\mathbf{v}_i)^T g(\mathbf{v}_j) \\ s_{ij}^{I \leftrightarrow M} = g(\mathbf{v}_i)^T h(\mathbf{w}_j, \mathbf{v}_j), \end{cases}$$

$$\text{Multimodal:} \begin{cases} s_{ij}^{M \leftrightarrow T} = h(\mathbf{w}_i, \mathbf{v}_i)^T f(\mathbf{w}_j) \\ s_{ij}^{M \leftrightarrow I} = h(\mathbf{w}_i, \mathbf{v}_i)^T g(\mathbf{v}_j) \\ s_{ij}^{M \leftrightarrow M} = h(\mathbf{w}_i, \mathbf{v}_i)^T h(\mathbf{w}_j, \mathbf{v}_j). \end{cases}$$

In total, there are 9 cross-modal combinations, resulting in 9 similarity matrices within each batch. We define our Omni-Retrieval loss as the sum of the contrastive loss over the 9 similarities:

$$\mathcal{L}_{omni} = \sum_{u,v \in \{I,T,M\}} \sum_i \delta_i^u \big[\log \frac{\exp(s_{ii}^{u \leftrightarrow v}/\tau)}{\sum_j \exp(s_{ij}^{u \leftrightarrow v}/\tau)} \\ + \log \frac{\exp(s_{ii}^{u \leftrightarrow v}/\tau)}{\sum_j \exp(s_{ji}^{u \leftrightarrow v}/\tau)} \big], \qquad (3)$$

where u and v refers to the modality of image, text, or multimodal, and δ_i^u indicates if the modality u of the i-th input pair exists or not, which works like a gate function to turn on/off the contrastive learning for each pair.

4 EXPERIMENTS

We conduct our pre-training in an efficient round-robin manner with only two types of forward/backward. At each step, we randomly choose the set of 5 image-text tasks or 9 Omni-Retrieval tasks for updating the model parameters. We evaluate the pre-trained model on 7 commerce-related tasks. We report experimental results on three model sizes: ResNet50-based 6-layer CommerceMM, ViT-B/16-based 6-layer CommerceMM, and ViT-B/16-based 12-layer CommerceMM We use MMF [33] for the implementation.

4.1 Pre-training Dataset

Our pre-training dataset consists of two subsets. The first subset is 102M image-text pairs, collected from two sources – Product Catalog Posts (50M) and Marketplace Posts (52M). The text is the concatenation of product title and description. The average number of words of the concatenated texts is 62. Note our texts are multilingual, including English, French, German, Spanish, Portuguese and other languages. For each image, we pre-compute its GrokNet embeddings [4] and ClusterFit probability [43] for the MIM tasks in our image-text pre-training. The second subset is 50M cross-modal and cross-pair data for the Omni-Retrieval pre-training. The data comes from various sources: (i) IG and FB Shops product catalog: Each product has product title/description and product image. (ii) IG and FB Shops text search queries with clicked product: The search query text is used as the query to retrieve the clicked products (product title/description, product image). (iii) IG and FB posts where a product is tagged on the post: The post usually contains a

post image and post caption text. The post caption text is frequently empty. Such posts are used to retrieve the tagged product (product title/description, product image) given the post image.

4.2 Downstream Tasks

With minimal surgery of CommerceMM, our model can be adapted to new tasks that are not seen during pre-training. We introduce 7 downstream commerce-related tasks. Each of the downstream fine-tuning tasks has its own dataset and annotations.

Catalog Categorization (CC) We annotated the fine-grained category labels for 2.5M multimodal catalog posts. The labels are tree-structured on shopping-based categories, e.g., "Home/Home Decor/Decorative Accents/Posters, Prints & Paintings". There are in total 5,168 leaf labels. We focus on classifying the leaf label for each post. One example of our catalog posts is shown in the product page of Fig. 1, consisting of an image, a title and a description. We concatenate the title and description as a full sentence. There are on average 99.6 words for each post. We add a 5,168-way classifier on top of the multimodal fusion encoder and apply cross entropy loss during the fine-tuning. Accuracy is used as the evaluation metric.

Marketplace Categorization (MPC) Similar to Catalog Categorization, we have also annotated the fine-grained category labels for 2M Marketplace posts. There are in total 1,387 leaf labels for the collected Marketplace posts. Each post consists of an image, a title and a description. We follow the same process concatenating the title and description. On average, there are 37.3 words in each text. We add a 1,387-way classifier on the multimodal fusion encoder's $\mathbf{w}^{cls}$ token for fine-tuning and inference. Overall accuracy is reported as the evaluation metric.

Image-Text Retrieval We collect 1M Catalog Posts for the text-to-image retrieval (**T2I**) and the image-to-text retrieval (**I2T**) tasks. We split a 10K subset for evaluation and leave the rest for model training. There are two ways to perform the image-text retrieval. One is using the ITM head to predict the matching score between the input image-text pair and rank the scores of all pairs [5, 11, 50]. One is computing the similarity between image encoder and text encoder and picks the best match without deeper fusion between the two modalities as [15, 19, 30]. While the first approach may achieve better performance [27], its computational cost is huge (quadratically increasing w.r.t the size of retrieval pool), which is not applicable in practice. Thus we follow the second approach to simply use our image and text encoders of CommerceMM for the retrieval tasks. We apply contrastive learning loss during the fine-tuning. Recall@1 is used to measure the retrieval performance.

Query-to-Product Retrieval (Q2P) We collected 1M text search queries with their clicked products. Each product is associated with a multimodal product page. We split a 10K subset for evaluation and leave the rest for model training. We consider this problem as a text-to-multimodal retrieval task. The model is shown in Fig. 3(a), where we use the CommerceMM's text encoder to encode the input query text and whole model to encode the candidate multimodal post. Contrastive learning loss is applied during fine-tuning. We use Recall@1 as the evaluation metric.

Image-Product Retrieval We collected 1M post image queries with the tagged products. Each pair consists of a source image and a target product page. We split out a 10K subset for evaluation and set

	Pre-training Tasks	Meta Avg.	CC	MPC	T2I	I2T	Q2P	I2P	I2P^i
1	None	50.39	72.08	63.75	22.70	23.88	48.43	55.80	66.10
2	MLM	52.77	73.10	67.94	24.87	25.84	51.59	60.46	65.59
3	MIM-kl	53.59	73.26	69.04	26.31	26.91	53.89	59.18	66.54
4	MIM-kl + MIM-fr	54.18	73.27	69.12	27.88	28.61	54.05	59.48	66.83
5	MLM + MIM-kl + MIM-fr	54.19	73.64	69.55	26.66	26.98	53.47	61.66	67.30
6	MLM + MIM-kl + MIM-fr + ITM	54.62	73.64	69.45	27.82	28.33	54.63	61.63	66.87
7	MLM + MIM-kl + MIM-fr + ITM + ITC	57.87	73.76	69.61	39.11	40.30	55.60	60.03	66.65
8	Omni Retrieval (Omni)	56.27	72.98	67.81	29.69	30.78	57.34	67.98	67.31
9	MLM + MIM-kl + MIM-fr + ITM + ITC + Omni	60.64	73.77	69.73	42.05	43.06	58.48	69.20	68.16

Table 1: Ablation Study of different pre-training tasks using the fine-tuned performance of Catalog Categorization (CC), Marketplace Categorization (MPC), Text-to-Image Retrieval (T2I), Image-to-Text Retrieval (I2T), Query-to-Product Retrieval (Q2P), Image-to-Product Retrieval (I2P), and Image-to-Product-Image Retrieval (I2P^i). All results are obtained from ResNet50-based CommerceMM (6-layer). For all retrieval tasks, R@1 scores are reported. Meta Avg is the average score of 7 downstream tasks, measuring the overall performance. Dark and light grey colors highlight the top and second best results for each task.
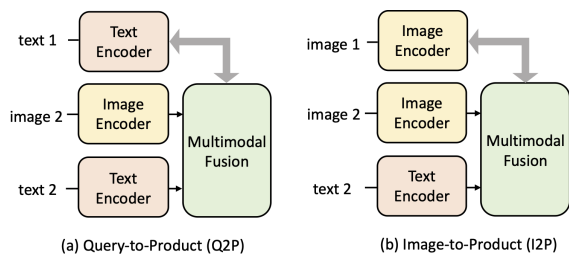

Figure 3: The model architecture for the Query-to-Product and Image-to-Product retrieval tasks.

up two tasks for this dataset. The first is the image-to-multimodal retrieval as the given data format, which is called image-to-product retrieval (**I2P**). The model for I2P is shown in Fig. 3(b), where we use the image encoder to encode the input query image and the whole model to encode the multimodal product post. The second is the image-to-image retrieval task, where we only use the target image as the candidates. We call this task as image-to-product-image retrieval (**I2P^i**). For both tasks, we apply contrastive learning loss and use Recall@1 for evaluation.

4.3 Ablation Study on Pre-training

We analyze the effectiveness of different pre-training settings through ablation studies over the 7 downstream tasks. In addition to the above mentioned standard metrics, we also compute the Meta Average score (average of the results across all tasks) as a global metric. The complete ablation study is listed in Table. 1. In these experiments, we use a ResNet50-based 6-layer CommerceMM, *i.e.*, the number of layers of text encoder and multimodal fusion together equals 6, initialized from XLM-R-small [6].

First, we provide a baseline without any multimodal pre-training involved in Line 1 (L1). In other words, this model is directly initialized from the off-the-shelf ResNet and XLM-R directly, which were pre-trained in vision-only and language-only domain.

Second, we validate the effectiveness of each pre-training task through a thorough ablation. Comparing L2 and L3, our proposed

MIM-kl (L3) achieves a clear gain over MLM (L2). When applying both MIM-kl and MIM-fr, L4 further improves the performance. This is a quite different observation compared with [10, 16], where previously MIM was not shown to be helpful. The difference indicates the effectiveness of our proposed MIM tasks. Interestingly, the combined MLM, MIM-kl, and MIM-fr in L5 does not quite outperform L4 (with only 0.01 gain on the Meta Average). One possible reason might be the two MIM tasks overshadows the effect of MLM. L6 and L7 adds ITM and ITC into the pre-training, both of which further improves the meta average. Notably, ITC introduces a significant improvement on image-text retrieval tasks, *i.e.*, T2I and I2T due to the same task has been well warmed up during pre-training.

Next, we validate the contribution of Omni-Retrieval (Omni) pre-training tasks. In L8, we only apply the 9 Omni-Retrieval tasks during pre-training without any help from image-text pre-training. We observe a significant gain on Q2P, I2P, and I2P^i, which benefits from the text-to-multimodal, image-to-multimodal, and image-to-image tasks in the Omni. Additionally, Omni also helps on the first 4 image-text tasks (CC, MPC, T2I, I2T), comparing L1 and L8. When combining both image-text and Omni-Retrieval pre-training, our model in L9 achieves the best across every single task.

Last but not least, we compare the performance between I2P and I2P^i. The two tasks share the same evaluation set, where each image in I2P^i is from its corresponding product page in I2P. Interestingly, we found that without Omni Retrieval, I2P^i always performs better than I2P in L1-L7. As comparison, L8 and L9 shows I2P's results are better than I2P^i with Omni, which aligns with our intuition that the multimodal product page contains more cues than its product image only. This indicates Omni helps learning more generalized representations under different alignment space.

4.4 Effectiveness of Vision Encoder, Text Encoder, and Model Size

In Table 2, we first show the results using the off-the-shelf GrokNet [4] Hash feature as the image embedding. We feed the text tokens and image embedding directly into the multimodal transformer as [5, 21], *i.e.*, the multimodal model is an early-fusion model without text encoder. We apply MLM, MIM-kl, MIM-fr and ITM for

	Vis. Enc.	(K, M)	Pre-training Tasks	Meta Avg.	CC	MPC	T2I	I2T	Q2P	I2P	I2P[i]
1	GrokNet Hash	(0, 6)	None	-	71.26	65.26	-	-	-	-	-
2	GrokNet Hash	(0, 6)	MLM + MIM-kl + MIM-fr + ITM	-	73.40	69.04	-	-	-	-	-
3	ResNet50	(0, 6)	MLM + MIM-kl + MIM-fr + ITM	-	73.54	69.13	-	-	-	-	-
4	ResNet50	(3, 3)	MLM + MIM-kl + MIM-fr + ITM	54.62	73.64	69.45	27.82	28.33	54.63	61.63	66.87
5	ResNet50	(3, 3)	MLM + MIM-kl + MIM-fr + ITM + ITC + Omni	60.64	73.77	69.73	42.05	43.06	58.48	69.20	68.16
6	ViT-B/16	(3, 3)	MLM + MIM-kl + MIM-fr + ITM + ITC + Omni	62.69	73.78	69.80	43.84	44.29	61.43	73.41	72.31
7	ViT-B/16	(6, 6)	MLM + MIM-kl + MIM-fr + ITM + ITC + Omni	66.85	74.31	70.60	52.10	53.72	65.70	77.16	74.36

Table 2: Effect of vision encoder, text encoder, and model size. (K, M) stands for (#text layers, #multimodal layers) inside CommerceMM, both of which are transformers. Dark and light grey colors highlight the top and second best results.

the multimodal pre-training, each of which is same as in Sec. 3.2[4]. The mere difference lies in the masking strategy of MIM, where we randomly shuffle the 0/1 bits of the input Hash for masking. For fine-tuning, we focused on the multimodal categorization tasks of CC and MPC. We observe that even with only 4 pre-training tasks, there is a significant improvement on CC and MPC tasks from L1 over L2, indicating that our pre-training also works well with a fixed vision encoder of GrokNet. We launched L2 as our current production model. More details of the deployment and product impact are provided in Sec. 4.8. We then show the advantage of end-to-end training. Comparing L2 and L3, we observe the end-to-end trained model with ResNet50 already outperforms GrokNet's Hash [4], which is from a ResNeXt101 model.

Next, we compare with and without text encoder. Both L3 and L4 have the same number of transformer layers, *i.e.*, the model size is the same, and are pre-trained with the same tasks. We observe that with a 3-layer text encoder, the model in L4 achieves better performance than the early-fusion model in L3 on CC and MPC. Moreover, the introduction of a text encoder allows us to perform those text-based retrieval tasks, *e.g.*, T2I, I2T, and Q2P. Adding the Omni-Retrieval pre-training tasks, the model in L5 further improves the performance of all the retrieval tasks, compared with L4.

We then compare the effectiveness of vision encoder. With the same input image size 224x224 and the same transformer architecture, we found ViT-B/16 [9] (L6) brings a consistent gain over ResNet50 (L5) on each of the 7 downstream tasks, showing a better visual encoder is beneficial to the multimodal applications.

We also experiment scaling up our transformer from 6 layers (L6) to 12 layers (L7). Note for the 12-layer transformer, we simply assign its first 6 layers from XLM-R to the text encoder and leave the rest 6 layers to the multimodal fusion encoder. Comparing L6 and L7, we observe a further improvement on each task with the larger model. We leave the exploration of even larger transformer of 24 or 36 layers to the future work.

4.5 Transferability to Academic Dataset

We also evaluate how our pre-trained model performs on the academic dataset, *e.g.*, FashionGen [31]. We strictly follow [50] constructing its image-text retrieval task. In its text-to-image retrieval, the model is required to pick the matched image from 101 images given a text. In the 101 images, one is positively paired with the text and the other 100 are randomly paired but sharing the same

sub-category as the positive, increasing the difficulty. The same setting is for its image-to-text retrieval. We fine-tune our smallest ResNet50-based 6-layer CommerceMM (L9 in Table 1) on the dataset with contrastive learning. In Table. 3, we compare our model with the state-of-art commerce-domain pre-trained models [11, 50]. We found even our smallest model already outperforms [11, 50] with a clear margin, indicating CommerceMM's superior transferability.

	FashionGen T2I			FashionGen I2T		
	R1	R5	R10	R1	R5	R10
FashionBERT [11]	26.8	46.5	55.7	24.0	46.3	52.1
KaleidoBERT [50]	33.9	60.6	68.6	28.0	60.1	68.4
CommerceMM (small)	39.6	61.5	72.7	41.6	64.0	72.8

Table 3: Image-Text Retrieval on FashionGen [31].

4.6 Modality Randomization

As in Sec. 3.1, our model design allows us to dynamically change the text encoder and multimodal fusion encoder by assigning different K and M layers to each. Previous works [1, 38, 39, 44] show the modal-agnostic training can be beneficial to the multimodal understanding. Our approach follows the same spirit. At each training step, we randomly assign K and M transformer layers (while keeping the sum of K and M unchanged) to the text encoder and multimodal encoder, so that every layer can share the knowledge from both text and multimodal modalities.

To validate this interesting idea, we conduct the modality randomization with the 5 image-text pre-training tasks. We then fine-tune our model on CC and MPC, both measuring the multimodal recognition capability. Fig. 4 compares the fixed-architecture and modality-randomized pre-trained models. Typically the pre-training and fine-tuning models should share the same architecture, but we propose to change the model architecture at fine-tuning stage for potential better performance. We observe the modality-randomized pre-training brings better performance than fixed-arch under any architecture setting at fine-tuning. We also found a smaller accuracy variance of the modality-randomized model under different architectures, showing its robustness to the model change.

We also experiment with changing the total number of layers. In Figure 5, we plot the CC and MPC accuracy with $K = 0$ text encoder and different M fusion layers (each is an early fusion model without text encoder). While it seems no surprise the deeper model brings better accuracy, we found our small models also perform well without notable performance drop compared with using full layers. Note our 2-layer model achieves 73.10 on CC and 67.88 on

---

[4]We cannot apply image-text contrastive learning or omni retrieval as there is no text encoder in the early-fusion architecture.
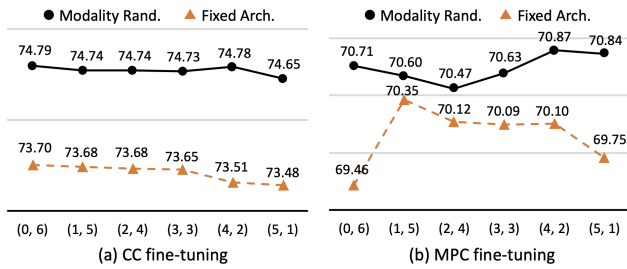
Figure 4: Effect of (K, M) from a fixed-architecture (3, 3) pre-trained model and modality-randomized pre-trained model. (K, M) stands for (#text layers, #multimodal layers).
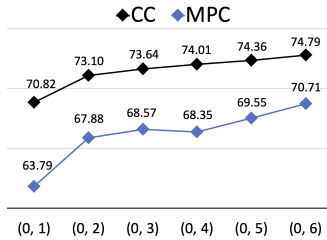

Figure 5: Effect of total number of layers from a modality-randomized model on CC and MPC. (K, M) stands for (#text layers, #multimodal layers).

MPC, which are already better than the 6-layer model without pre-training in Table 1 L1. Thus with modality randomization, we can flexibly adjust our model architecture to satisfy different efficiency constraints in practice.

Modality randomization can be further explored by incorporating it into the proposed pre-training in 1 stage. We leave this promising idea to future work.

4.7 Visualization

We visualize the embeddings from the image encoder, text encoder, and multimodal fusion encoder respectively in Fig. 6. Specifically, we feed 1K multimodal catalog posts from 10 most popular categories into CommerceMM. T-SNE is applied for the visualization and the colors are corresponding to the annotated categories. We compare the results of without pre-training (initialized from ResNet+XLM-R), image-text pre-training with MLM, MIM-kl, and MIM-fr, and full 14-task pre-training. We observe that with additional retrieval tasks in pre-training, *e.g.*, ITC and Omni, the embeddings of the same class are better clustered with closer distance. This indicates our proposed pre-training tasks helps learning a more discriminative representation for each modality.

4.8 Deployment and Product Impacts

An early version of CommerceMM model has been deployed at Meta, *i.e.*, L2 in Table 2. We pre-trained the model using a subset of the tasks (MLM, MIM-kl, MIM-fr and ITM) and then fine-tuned on three downstream tasks. We confirmed the benefits of this framework by running A/B test experiments on Catalog Categorization
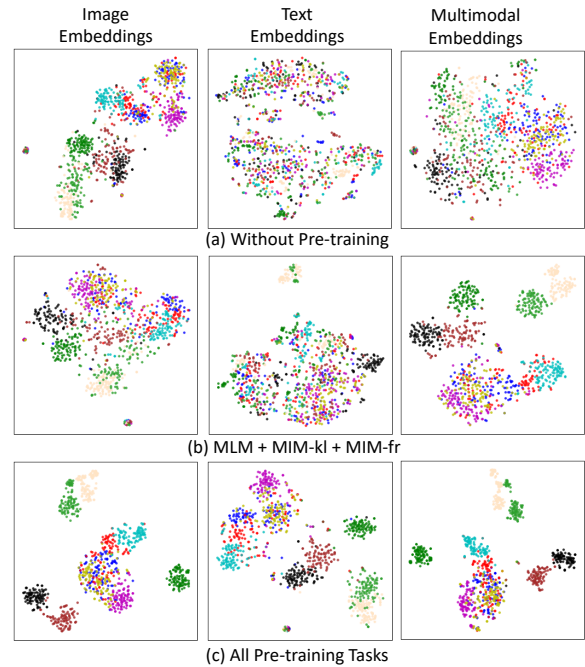

Figure 6: T-SNE of different pre-trained models' embeddings.

(CC), Marketplace Categorization (MPC) and Catalog Attributes applications.

Catalog Category Filters Users on Shops have access to category filters which lets them search products for a specific category within a shop. With our newly launched model, we have increased the number of shops with category filter by over 4 times.

Catalog Attributes Filters Attributes are fine-grained characteristic of objects, such as color, pattern, material, etc. We ran A/B tests to confirm the improvement of attributes prediction on IG and FB Shops for three specific product attributes - color, gender and material. This model enabled a launch of these three attribute filters on 52.7% of all Shops on the platform due to significant precision and coverage improvements compared to the baseline production models. After inference, all these category and attributes predictions are stored in a distributed key-value store, which gets consumed by multiple product groups across the company to improve various ranking and recommendation products.

This early deployed version gives a substantial evidence that the pre-training is helpful for downstream tasks. The end-to-end CommerceMM will be deployed in the same way with a simplification of removing the service call of GrokNet Hash [4], *i.e.*, the inference will be more efficient.

5 CONCLUSION

We introduce CommerceMM, a large-scale commerce multimodal model at Meta. We present 5 effective image-text pre-training tasks, and propose a novel set of 9 Omni-Retrieval tasks on cross-modal and cross-pair data. Pre-trained on the large-scale diverse multilingual multimodal data, CommerceMM outperforms state-of-art models across 7 downstream tasks with a clear margin.

REFERENCES

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *NeurIPS* (2021).

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *ICCV*.

[4] Sean Bell, Yiqun Liu, Sami Alsheikh, Yina Tang, Edward Pizzi, M Henning, Karun Singh, Omkar Parkhi, and Fedor Borisyuk. 2020. Groknet: Unified computer vision model trunk and embeddings for commerce. In *KDD*.

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Learning universal image-text representations. In *ECCV*.

[6] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

[8] Xiao Dong, Xunlin Zhan, Yangxin Wu, Yunchao Wei, Xiaoyong Wei, Minlong Lu, and Xiaodan Liang. 2021. M5product: A multi-modal pretraining benchmark for e-commercial product downstream tasks. *arXiv preprint arXiv:2109.04275* (2021).

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[10] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Zicheng Liu, Michael Zeng, et al. 2021. An Empirical Study of Training End-to-End Vision-and-Language Transformers. *arXiv preprint arXiv:2111.02387* (2021).

[11] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. 2020. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *SIGIR*.

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2021. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377* (2021).

[13] Xiaowei Hu, Xi Yin, Kevin Lin, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. 2021. Vivo: Surpassing human performance in novel object captioning with visual vocabulary pre-training. In *AAAI*.

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

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.

[16] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*.

[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. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV* (2017).

[18] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*.

[19] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS* (2021).

[20] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *EMNLP*.

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

[22] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2020. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409* (2020).

[23] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

[25] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS* (2019).

[26] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. 2018. Exploring the limits of weakly supervised pretraining. In *ECCV*.

[27] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2021. Thinking Fast and Slow: Efficient Text-to-Visual Retrieval with Transformers. In *CVPR*.

[28] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *NeurIPS* (2011).

[29] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

[31] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. 2018. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317* (2018).

[32] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

[33] Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2020. MMF: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf.

[34] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530* (2019).

[35] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *ICCV*.

[36] Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS* (2017).

[38] Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2021. UFO: A UniFied TransfOrmer for Vision-Language Representation Learning. *arXiv preprint arXiv:2111.10023* (2021).

[39] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. 2021. VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. *arXiv preprint arXiv:2111.02358* (2021).

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

[41] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. 2021. Masked Feature Prediction for Self-Supervised Visual Pre-Training. *arXiv preprint arXiv:2112.09133* (2021).

[42] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2021. Simmim: A simple framework for masked image modeling. *arXiv preprint arXiv:2111.09886* (2021).

[43] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. 2020. Clusterfit: Improving generalization of visual representations. In *CVPR*.

[44] Haoxuan You, Luowei Zhou, Bin Xiao, Noel C Codella, Yu Cheng, Ruochen Xu, Shih-Fu Chang, and Lu Yuan. 2021. MA-CLIP: Towards Modality-Agnostic Contrastive Language-Image Pre-training. (2021).

[45] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934* (2020).

[46] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *CVPR*.

[47] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*.

[48] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *AAAI*.

[49] Yushan Zhu, Huaixiao Zhao, Wen Zhang, Ganqiang Ye, Hui Chen, Ningyu Zhang, and Huajun Chen. 2021. Knowledge perceived multi-modal pretraining in e-commerce. In *ACM-MM*.

[50] Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. 2021. Kaleido-bert: Vision-language pre-training on fashion domain. In *CVPR*.

A   APPENDIX

A.1   Training Details

During pre-training, we set the learning rate as 5e-5, batch size as 2,048, and update in total 300K steps. We apply another 100K steps for a 2nd-stage pre-training when modality randomization is applied. In the 5 image-text pre-training tasks, we assign equal weights to MIM-kl, MIM-fr, ITM and ITC and half of their weights to MLM. We found equal weight on MLM will hurt the overall performance. One possible reason is MLM's effectivenss is overshadowed by the more effective MIM tasks. For the 9 Omni-Retrieval tasks, we also assign the same weight to each of them. Our model is trained in a round-robin fashion. At each step, we randomly pick between image-text pre-training and Omni-Retrieval pre-training for updating the model parameters. It takes in total 3,840 and 4,608 A100 hours to end-to-end pre-train our 6-layer CommerceMM and 12-layer CommerceMM with the vision encoder respectively.

A.2   More Visualization

We also visualize the text-to-image attention in Fig. 7. Comparing the the models pre-trained with 5 image-text tasks and all tasks (with Omni-Retrieval pre-training), we observe the cross-modal attention from all pre-training can better attend to the right regions referred by the key words, *e.g.*, "dress" and "earrings".
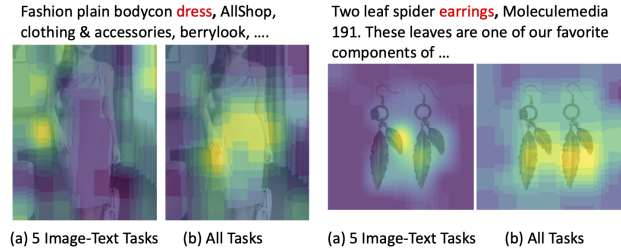

Figure 7: Visualization of text-to-image attention from different pre-trained models.