M5Product: Self-harmonized Contrastive Learning for E-commercial Multi-modal Pretraining

Xiao Dong[1†], Xunlin Zhan[1,2†], Yangxin Wu[1], Yunchao Wei[3], Michael C. Kampffmeyer[4], Xiaoyong Wei[5], Minlong Lu[6], Yaowei Wang[5], Xiaodan Liang[1,2⋆]

[1]Sun Yat-sen University [2]Shenzhen Campus of Sun Yat-sen University [3]Beijing Jiaotong University
[4]UiT The Arctic University of Norway [5]PengCheng Laboratory [6]Alibaba Group.

{dongx55, zhanxlin, wuyx29}@mail2.sysu.edu.cn, {dx.icandoit,wychao1987,xdliang328}@gmail.com,
ymlml@zju.edu.cn, michael.c.kampffmeyer@uit.no, cswei@scu.edu.cn, wangyw@pcl.ac.cn

Abstract

*Despite the potential of multi-modal pre-training to learn highly discriminative feature representations from complementary data modalities, current progress is being slowed by the lack of large-scale modality-diverse datasets. By leveraging the natural suitability of E-commerce, where different modalities capture complementary semantic information, we contribute a large-scale multi-modal pre-training dataset **M5**Product. The dataset comprises 5 modalities (image, text, table, video, and audio), covers over 6,000 categories and 5,000 attributes, and is 500× larger than the largest publicly available dataset with a similar number of modalities. Furthermore, **M5**Product contains incomplete modality pairs and noise while also having a long-tailed distribution, resembling most real-world problems. We further propose **S**elf-harmonized **C**ontr**A**stive **LE**arning (**SCALE**), a novel pretraining framework that integrates the different modalities into a unified model through an adaptive feature fusion mechanism, where the importance of each modality is learned directly from the modality embeddings and impacts the inter-modality contrastive learning and masked tasks within a multi-modal transformer model. We evaluate the current multi-modal pre-training state-of-the-art approaches and benchmark their ability to learn from unlabeled data when faced with the large number of modalities in the **M5**Product dataset. We conduct extensive experiments on four downstream tasks and demonstrate the superiority of our **SCALE** model, providing insights into the importance of dataset scale and diversity. Dataset and codes are available at [1].*

1. Introduction

Self-supervised learning has been driving the rapid development of fields such as computer vision and natural language processing, as well as research on multi-modal representation learning. In particular, it has been shown both from a theoretical [18] and a practical [16, 58] perspective that large scale datasets with **diverse modalities** can effectively enhance the discrimination of generated features and thus improve the performance in vision-language tasks. However, current advances are severely limited by the lack of such large-scale diverse-modality datasets, with the largest public multi-modal datasets only containing text and image modalities and no category information [41].

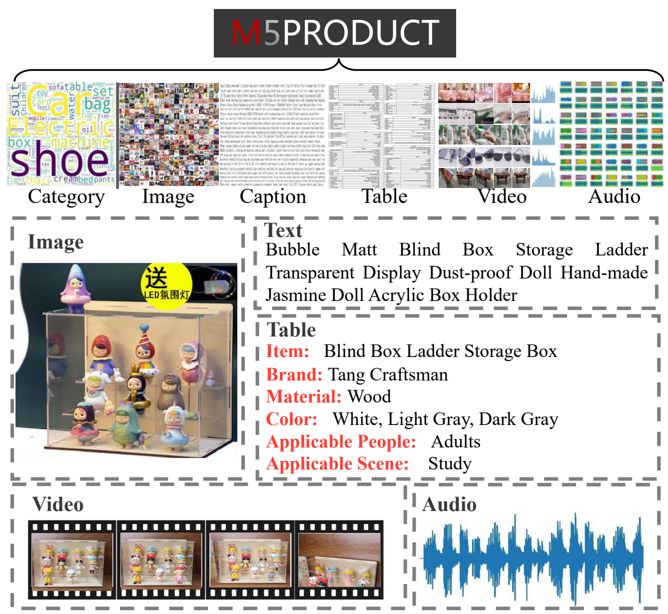


Figure 1. Our **M5**Product dataset contains a large variety of modalities (image, text, table, video and audio) that depict the categories, descriptions, materials, properties and purposes of E-commerce products, and diverse real-world data samples.

Given the prevalence of online shopping in daily life, with its natural occurrence of multi-modal information

[1] https://xiaodongsuper.github.io/M5Product_dataset/
† Equal contribution. ⋆ Corresponding Author.

and diverse categories, multi-modal pre-training on E-commercial products has received increasing attention and led the developments of next-generation technology for several downstream tasks (e.g., multi-modal retrieval, multi-modal classification, and clustering). However, even among the present product datasets (e.g., RPC checkout [48], Dress Retrieval [9] and Product1M [55]), the number of categories is insufficient to robustly verify the performance of downstream tasks.

More importantly, the current research community mostly focuses on two modalities (text and image) in both general multi-modal and E-commerce datasets, while ignoring the importance of additional complementary information from structural data as well as video and audio modalities. Tabular data for instance can provide detailed information about properties and characteristics, such as brand, materials, attributes, and scenarios, while audio and video can convey different perspectives, scales, affordances, selling points, characteristics, and use scenarios that are not obvious from images or text alone. The focus on these two modalities is partly due to the lack of datasets with diverse modalities as well as an under-exploration of approaches to balance the modality importance in these settings. In particular, two key challenges are: 1) Modality Interaction: How to learn common representations from unimodal, bimodal, trimodal, and even multi-modal relationships between different modalities using an elegant approach that scales to a large number of modalities; 2) Modality Noise: How to reduce the influence of modality noise (missing and incompleted modalties) during the training process.

To address the problem of insufficient modality diversity and limited scale, while at the same time providing a challenging real-world scenario, we present a very large-scale E-commerce multi-modal product dataset **M5**Product, which is one of the largest and most diverse multi-modal product dataset to date. Our **M5**Product dataset contains more than 6 million multi-modal samples from 6,232 categories and has more complex and diverse modalities than existing datasets. This allows **M5**Product to be used for more comprehensive evaluation of the practical application and generalization abilities of multi-modal pretraining models and can improve the modality fusion performance, facilitating new directions in multimodal research. Figure 1 shows the five modalities (image, caption, video, audio, and specification (table)) of our dataset.

To further address the modality fusion limitations of existing methods as well as handle modality noise, we propose a generic framework that takes five-modality data as inputs, as shown in Figure 2. The framework consists of a simple and efficient multi-modal five stream pre-training model named **S**elf-harmonized **C**ontr**A**stive **LE**arning (**SCALE**) and is evaluated on several downstream tasks and compared with several recent state-of-the-art vision-language models [7, 27, 30, 38, 42, 45, 56]. **SCALE** increases modality alignment effectiveness by implementing a self-harmonized strategy that adapts the alignment weights between different modalities in the contrastive learning modules and masked tasks to adaptively integrate complementary modality information. In summary, our contributions are as follows:

- We provide the largest five-modality E-commerce dataset **M5Product**. Through its large scale, diversity, complex real scenarios and number of modalities, **M5Product** provides a comprehensive environment for evaluating the generalization performance of multi-modal pre-training models.
- Our Self-harmonized Contrastive Learning (**SCALE**) framework learns adaptive modality interactions, resulting in more effective modality fusion. We compare **SCALE** to a comprehensive set of baseline methods and demonstrate its superior performance on the M5Product dataset.
- *Interesting Observations*: 1) In large-scale and complex scenarios, the complementary gain of different modalities increases. Learning modality alignment weights allows our **SCALE** framework to effectively coordinate complementary information to achieve better performance. 2) For multi-modal pre-training models in the E-commerce domain, dataset scale and diversity are relatively important for the downstream tasks. Given the large-scale and diverse products, our **SCALE** framework generalizes better than other baselines to downstream tasks.

2. Related Work

Multi-modal pre-training datasets. Most multi-modal pre-training datasets are collected from social websites (e.g., Twitter and Facebook) and are limited to just two modalities collected for specified tasks. These datasets can be divided into four categories according to their modality composition, i.e., audio/text, video/text, image/text, and others. Among these, LJ Speech [19] and SQuAD [25] are classical audio/text datasets and used for voice synthesis and audio Q&A, while most video/text datasets [2, 20, 24, 32, 46, 47, 51, 57] are used for video Q&A. However, these datasets commonly only contain a limited number of samples, limiting their applicability to multi-modal pretraining. Image/text datasets [1, 4, 8, 17, 22, 23, 29, 34, 41, 43, 48, 53], on the other hand, tend to be larger and have been widely used for pretraining multi-modal models. Among these, the CC 3M [41] with more than three million image-text pairs is the most widely used pre-training dataset, and has recently been expanded to CC 12M [5], the largest text-image cross-modal dataset currently. Apart from these, commonly used Image/text datasets for multi-modal retrieval tasks are MS COCO [29], Flickr30K [53], INRIA-Websearch [22] and

Table 1. Comparisons with other widely used multi-modal datasets. "-" means not mentioned. Our **M5**Product is one of the largest multi-modal datasets compared with existing datasets. Six modalities are separately denoted as: Image (I), Text (T), Video (V), Audio (A), Table (Tab) and 3D Image (3D).

Dataset	Samples	Categories	Instances	Modalities	Modal type	Product
SQuAD [25]	37,111	-	-	2	A/T	no
HowTo100M [32]	1,220,000	12	-	2	V/T	no
CC 3M [41]	3,300,000	-	-	2	I/T	no
CC 12M [5]	12,423,374	-	-	2	I/T	no
CMU-MOSEI [54]	23,500	2	-	3	T/V/A	no
XMedia [36]	12,000	20	-	5	I/T/V/A/3D	no
RPC checkout [48]	30,000	200	367,935	2	I/T	yes
Dress Retrieval [9]	20,200	50	∼20,200	2	I/T	yes
Product1M [55]	1,182,083	458	92,200	2	I/T	yes
MEP-3M [6]	3,012,959	599	-	2	I/T	yes
M5Product	**6,313,067**	**6,232**	-	**5**	**I/T/V/A/Tab**	**yes**

NUS-WIDE [8] with standard annotations. Other datasets include CMU-MOSEI [54] and XMedia [36], where CMU-MOSEI mainly focuses on the emotional analysis and XMedia is utilized for cross-modal retrieval.

Aside from the abovementioned datasets, there exist several E-commerce datasets. The Dress Retrieval [9], RPC checkout [48] and Product1M [55] are typical E-commerce multi-modal datasets. The Dress Retrieval dataset contains 20,200 samples from 50 clothing categories, RPC checkout offers 30,000 samples of small retail goods on simple backgrounds and Product1M provides 1.18 million samples from 458 cosmetics classes. Compared with these three datasets, our **M5**Product is not only larger in terms of categories and data scale, but also contains a more diverse set of modalities. A detailed comparison with other multi-modal pre-training datasets is provided in Table 1.

Multi-modal pre-training for E-commerce products. Several vision-language pre-training models have been explored for visual-text multi-modal learning in recent years. They can coarsely be divided into two categories: 1) Single-stream models whose transformer layer operates collectively on the concatenation of the visual and text inputs, e.g, VL-bert [42], Image-BERT [37], VideoBERT [44], MMT [12], HERO [26], VisualBERT [27] and UNITER [7]. 2) Dual-stream models whose image and text inputs are not concatenated, such as ViLBERT [30], LXMERT [45], CLIP [38] and DALL-E [39].

Within E-commerce, fashion-based tasks have been addressed in among others FashionBERT [13], MAAF [11], Kaleido-BERT [59], M6 [28] and CAPTURE [55]. All existing studies in the E-commerce scenarios focus solely on the image and text modalities and none of the approaches can utilize more modalities. Besides, all existing methods default to assigning the same contribution to different modalities when modeling multi-modal interactions.More specifically, transformer-based approaches combine high-level features that are extracted from the different inputs via concatenation, where the uni-modal transformers are

Table 2. The characteristics of different modalities for E-products.

Modality	APP	USA	SPEC	SELL	PROD	MATE	CATE
Image	✓						
Text		✓		✓	✓		✓
Video	✓	✓			✓	✓	
Audio		✓			✓	✓	
Table			✓		✓	✓	✓

trained via masked task constraints or via constructing inter-modality losses between different modalities. This restricts the models from effectively prioritizing modalities and tends to limit performance improvements as the number of modalities increases.

Our proposed benchmark fills this gap by exploiting all the diverse modalities of the **M5**Product dataset and provides a strong baseline for multi-modal pre-training research in the field of E-commerce and beyond.

3. M5Product Dataset

Data Collections. The dataset is crawled from a popular E-commerce website [2]. and the front page of each E-commerce product is analyzed to collect the multi-modal information consisting of product images, captions, videos, and specifications (table information) [3]. Duplicate data is removed and audio information is extracted from videos via the **moviepy** [4] tool and saved in mp3 format. For product specifications, we extract 5,679 product properties and 24,398,673 values to construct a table database coarsely labeled by e-commerce merchants. After processing, the dataset contains 6,313,067 samples. Note, being a real-world dataset, our **M5**Product is, unlike traditional multi-modal datasets, not a complete paired dataset and contains samples with only a subset of modalities as well as long-tailed distributed (Figure 3). We summarize the product characteristics that are relayed by the different modalities in our dataset in Table 2, where APP, USA, SPEC, SELL, PROD, MATE and CATE denote Appearance, Usage, Specification, Selling Point, Production, Material and Category Descriptions, respectively.

Quantitative analysis. 1) **Diversity**: The dataset consists of more than 6,000 classes covering various and massive amounts of E-commerce products such as clothes, cosmetics, and instruments. Figure 1 illustrates the diversity of the modalities and categories and we further provide a description of the data format and the collection process in Section E of the supplementary materials. Finally, a quantitative analysis of the category and modality distributions can be found in Section F. Note that about 5% of products are unimodal samples *e.g.* only either contain images, captions,

[2] We are authorised by the company to access and obtain the data. We are further authorised to share the dataset and the detailed license is given in Section A of the supplementary material [3] In this work we focus on core data modalities (image, text, video, audio, and table data) only and do not consider extracted feature representations such as OCR and Motion embeddings that are extracted from core modalities as separate modalities.

[4] https://pypi.org/project/moviepy/

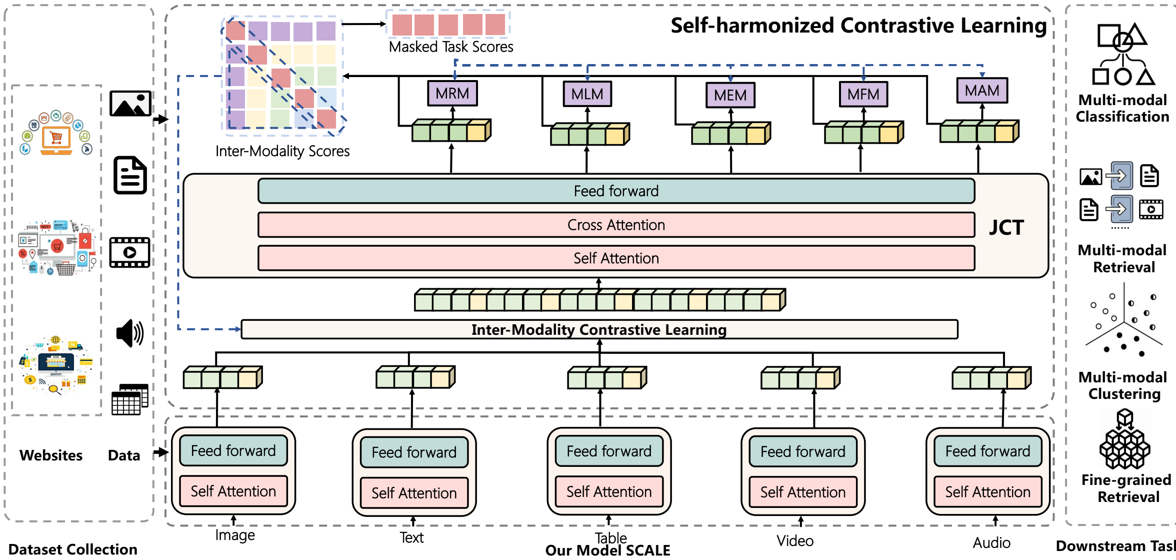


Figure 2. An illustration of our **M5**Product benchmark. It consists of a five-modality E-commerce dataset with a more diverse and complex backgrounds collected from the real-world online-shopping website. It also proposes a **SCALE** model to capture the maximum modality complementary information for four common downstream tasks: 1) multi-modal retrieval, 2) fine-grained retrieval, 3) multi-modal classification, and 4) multi-modal clustering. The benchmark verifies the effectiveness of modality diversity in five widely used modalities.

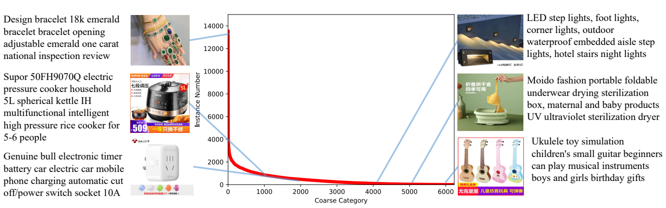


Figure 3. Training data distribution over whole categories.

or tabular properties. 2) **Quality**: We further provide a comparison between our **M5**Product dataset and some widely-used datasets for multi-modal pre-training in Table 1. A more extensive comparison with other multi-modal datasets can be found in Section H of the supplementary materials. Compared with existing multi-modal datasets, **M5**Product is the first extremely large public real-world E-commerce product dataset that contains data of more than two modalities.

Moreover, our dataset contains a large amount of instances, i.e., more than six million samples from the 6,232 coarse categories. These abundant data will benefit several downstream tasks such as self-learning, weakly-supervised learning, multi-modal retrieval, cross-modal generation and fine-grained recognition.

Additional analysis. In the supplementary materials, we provide dataset collection details in Section B and detail how the dataset is split into training and test in Section D and how annotations are obtained in Section C. We further provide a smaller split, referred to as *subset*, which is used to show the difference in performance for a smaller dataset. Finally, we provide further insights into the composition of the dataset (missing modalities, unimodal data analysis, and data format) in supplementary Section F.

4. Our Methodology

As shown in Figure 2, our **SCALE** framework consists of a self-harmonized contrastive learning module and a self-supervised multi-modal transformer. In this section, we first provide the architectural design of **SCALE** in Section 4.1 before describing the five masked tasks that enable the self-supervised learning of **SCALE** in Section 4.2. Finally, we present the detailed learning process of **SCALE** and detail how multi-modal alignment is achieved in Section 4.3.

4.1. Architectural Design of SCALE

As depicted in Figure 2, **SCALE** is a typical single-stream transformer architecture. In the bottom part, the Image/Text/Table/Video/Audio embedding layers and transformers aim to extract modality features and generate token features. Specifically, the text and table encoders are standard transformers to encode the caption and table information of products, respectively. The image encoder instead takes proposals extracted by bottom-up-attention [3] as inputs, while ordinal frames sampled from the video are fed into the video encoder. For the audio encoder, **SCALE** extracts MFCC [33] features from audio. After being processed by the separate modality encoders, the token features of different modalities are concatenated and fed into a Joint Co-Transformer (**JCT**) module to capture the token relationships between different modalities.

Missing Modalities. Zero imputation of missing modalities is leveraged to utilize all available data when training

SCALE. We provide experimental evidence that **SCALE** benefits from the incomplete samples in Section I of the supplementary material.

4.2. SCALE by Masked Multi-Modal Tasks

Similar to previous works, we utilize several **pre**text tasks (**PRE**) to facilitate self-supervised learning of **SCALE** in the Joint Co-Transformer module. For modality-wise feature learning from the image and text modalities, we adopt the Masked Region Prediction task (MRP) and the Masked Language Modeling task (MLM), respectively, after the **JCT**. Utilizing the characteristics of the table, video, and audio modalities, we further propose a Mask Entity Modeling task (MEM), Mask Frame Prediction task (MFP), and Mask Audio Modeling task (MAM) following a similar strategy of predicting masked tokens. In all masked tasks, the ground-truth labels are the features of masked areas. For all masking tasks, 15% of the inputs are masked out and the remaining inputs are used to reconstruct the masked information. Please note that unlike in the MLM task, where 15% of individual words are masked, 15% of the entities (properties, brand names, etc.) are entirely masked out for the MEM task. This drives our model to learn better table representations to recover the masked inputs, which is illustrated in Section 5.3. The loss function of the ith modality is defined as:

$$\mathcal{L}_{M_i}(\theta) = -E_{t_{msk}\sim\mathbf{t}} \log P_\theta\left(t_{msk} \mid t_{\neg msk}, \mathbf{M}_{\neg i}\right), \quad (1)$$

where $t_{\neg msk}$ denotes the unmasked tokens surrounding the masked token t_{msk}, θ represents the network parameters, and M_i and $M_{\neg i}$ are the ith modality and the remaining modalities, respectively.

4.3. Self-harmonized Inter-Modality Contrastive Learning

Self-harmonized Inter-Modality Contrastive Learning (**SIMCL**) is at the core of our proposed **SCALE** framework. It aims to facilitate the semantic alignment between different modalities via a self-harmonized strategy for adaptive **I**nter-**M**odality **C**ontrastive **L**earning (**IMCL**). For a minibatch of modality samples $D \in R^{B\times M\times F}$, where B, M and F denote the batch size, number of modalities, and embedding dimension, respectively, we first construct the contrastive loss between each modality.

Given N data samples $\{(d_i^{(0)}, d_i^{(1)})\}_{i=1}^{N}$, where each sample has two modalities (0) and (1), we select the N modality pairs as positive pairs in our contrastive learning. For each positive pair $(d_i^{(0)}, d_i^{(1)})$, negative pairs are constructed by pairing $d_i^{(0)}$ and $d_i^{(1)}$ with the remaining N-1 samples from the other modality, resulting in 2(N-1) negative pairs. For a modality pair $(d_i^{(0)}, d_i^{(1)})$ and their embedding features $(f_i^{(0)}, f_i^{(1)})$, the cross-modal contrastive loss of each modality pair is:

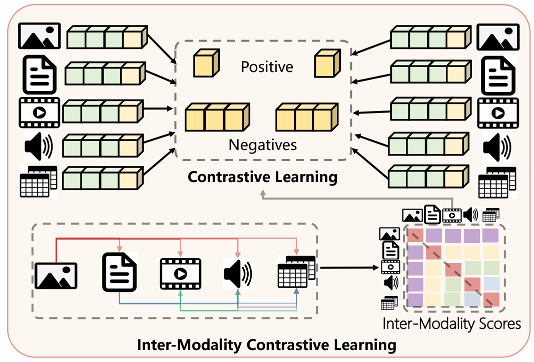


Figure 4. The Inter-Modality Contrastive Learning module of our **SCALE** framework.

$$\mathcal{L}_{CL}(d_i^{(0)}, d_i^{(1)}) = -\log \frac{\exp\left(\mathrm{sim}\left(f_i^{(0)}, f_i^{(1)}\right)/\tau\right)}{\sum\limits_{m=0}^{1}\sum\limits_{k=1}^{N} \mathbf{1}_{[k\neq i]} \exp\left(\mathrm{sim}\left(f_i^{(m)}, f_k^{(1-m)}\right)/\tau\right)}, \quad (2)$$

where sim is the cosine similarity, τ is the temperature parameter and $\mathbf{1}_{[k\neq i]}$ is a binary indicator function, and $\mathbf{1}$=1 for $k \neq i$ and 0 otherwise.

In most prior work, only two modalities are considered and Eq. 2 can be used. However, when considering trimodal data or data with even more than three modalities, it is not suitable to directly fit the loss function as it does not account for the difference in complementary information that different modalities contribute. To solve this problem, we define a simple but effective self-harmonized method to model the complementary process of the inter-modal relationships. We introduce a modality alignment score matrix, to encode the relationships among the inter-modal losses $\mathcal{L}_{CL}$ and the intra-modal losses $\mathcal{L}_{M_i}$. The alignment score matrix S for each data sample is initialized by a zero matrix and updated as free model parameters. To obtain modality importance scores for each modality combination, we apply the softmax function to S. Finally, the importance scores are multiplied to generate the modality alignment score S as $S = S \cdot \mathrm{softmax}(S)$. The learning process is shown in Figure 4 and illustrates that **SIMCL** takes full advantage of the inter-modal relationships. Given the modality alignment score S, the triangular part $S_\bigtriangledown$ is selected to weight the inter-modal losses $\mathcal{L}_{CL}$ and the diagonal part $S_\setminus$ is utilized to constrain the intra-modal losses $\mathcal{L}_{M_i}$, resulting in the weighted loss:

$$\mathcal{L}_{total} = \sum\nolimits_{S_{i,j}}^{S_\bigtriangledown} \mathcal{L}_{CL_{i,j}}(S_{i,j} logit_{i,j}) + \sum\nolimits_{S_i}^{S_\setminus} \mathcal{L}_{M_i}(S_i logit_i) \quad (3)$$

where logit is the loss logit.

5. Experiments

Implementation Details. We use BERT [10] to initialize the text transformer of our proposed **SCALE** framework,

Table 3. The (pretrain/finetune) performance gains from sequentially adding more modalities using **SCALE** on the subset (top) and the whole dataset (bottom). The retrieval performances are based on the features extracted from pretrain and finetune stages.

Modality	Accuracy	mAP@1	mAP@5	mAP@10	Prec@1	Prec@5	Prec@10
Text	77.42	47.70 / 65.10	53.63 / 68.39	51.59 / 66.99	47.70 / 65.10	30.96 / 44.89	24.15 / 33.44
+Image	79.58	51.47 / 67.02	56.16 / 69.85	54.41 / 68.43	51.47 / 67.02	33.41 / 46.29	25.55 / 34.29
+Table	82.83	57.14 / 67.97	61.71 / 70.34	59.64 / 69.38	57.14 / 67.97	38.02 / 46.85	28.99 / 34.36
+Video	84.31	58.57 / 69.79	63.15 / 72.30	61.02 / 70.67	58.57 / 69.79	39.26 / 47.44	29.56 / 34.78
+Audio	**85.50**	**58.72 / 70.62**	**63.17 / 73.02**	**61.05 / 71.50**	**58.72 / 70.62**	**39.66 / 48.20**	**30.32 / 35.35**
Text	81.11	55.82 / 69.47	60.74 / 72.74	59.02 / 71.79	55.82 / 69.47	36.99 / 48.76	28.04 / 35.84
+Image	83.68	59.81 / 71.51	64.13 / 74.51	62.18 / 73.21	59.81 / 71.51	38.97 / 49.27	30.15 / 36.72
+Table	84.63	61.32 / 72.34	65.53 / 74.86	63.62 / 73.47	61.32 / 72.34	40.66 / 49.77	30.78 / 36.95
+Video	84.90	62.65 / 72.59	65.67 / 75.05	63.87 / 73.62	62.65 / 72.59	41.18 / 49.96	31.01 / 37.04
+Audio	**86.57**	**63.56 / 73.77**	**67.51 / 76.17**	**65.39 / 74.73**	**63.56 / 74.01**	**42.68 / 50.78**	**32.17 / 37.42**

Table 4. The performance of our model **SCALE** under different modality combinations on the coarse- and fine-grained multi-modal retrieval and classification tasks. In the following, I, T, Tab, V and A denote image, text, table, video and audio modalities, respectively.

Modality Combinations	Accuracy	mAP@1	mAP@5	mAP@10	Prec@1	Prec@5	Prec@10
I+Tab	62.00	44.53 / 45.97	49.62 / 51.89	48.28 / 50.33	44.53 / 45.97	30.89 / 34.08	23.65 / 28.63
I+V	34.57	20.57 / 36.29	26.78 / 42.72	26.41 / 41.38	20.57 / 36.29	14.71 / 26.52	11.78 / 22.34
I+A	27.67	15.73 / 35.64	20.85 / 42.96	20.72 / 41.70	15.73 / 35.64	11.16 / 27.02	9.47 / 22.78
I+T	**79.58**	**67.02 / 62.20**	**69.85 / 66.97**	**68.43 / 64.21**	**67.02 / 62.20**	**46.29 / 49.85**	**34.29 / 42.36**
I+T+V	80.34	67.35 / 63.05	70.29 / 67.37	68.95 / 64.62	67.35 / 63.05	46.45 / 50.85	34.33 / 43.02
I+T+A	79.73	67.19 / 64.21	70.15 / 68.25	68.64 / 65.35	67.19 / 64.21	46.33 / 50.42	33.32 / 42.93
I+Tab+V	63.09	45.94 / 47.33	51.32 / 53.33	49.78 / 51.28	45.94 / 47.33	31.69 / 35.81	24.12 / 30.05
I+T+Tab	**82.83**	**67.97 / 68.30**	**70.34 / 72.67**	**69.38 / 70.07**	**67.97 / 68.30**	**46.85 / 57.44**	**34.36 / 50.59**
I+T+Tab+V	84.31	69.79 / 68.40	72.30 / 72.91	70.67 / 70.31	69.79 / 68.40	47.44 / 57.60	34.78 / 51.47
I+Tab+A+V	63.54	47.24 / 48.24	52.07 / 53.89	50.41 / 51.89	47.24 / 48.24	32.19 / 36.29	24.47 / 30.74
I+T+A+V	80.36	68.80 / 66.43	70.84 / 71.12	69.71 / 68.16	68.80 / 66.43	47.24 / 54.03	34.57 / 47.53
I+T+Tab+A	**84.33**	**70.23 / 68.97**	**72.59 / 73.07**	**70.94 / 70.77**	**70.23 / 68.97**	**47.58 / 57.89**	**35.33 / 51.60**
I+T+Tab+A+V	**85.50**	**70.62 / 69.25**	**73.02 / 74.08**	**71.50 / 71.02**	**70.62 / 69.25**	**48.20 / 58.76**	**35.35 / 52.05**

Table 5. Comparisons of image and text modalities on the subset (top) and the whole dataset (bottom).

Method	mAP@1	Accuracy	NMI	Purity
Image_{based}	15.17	27.67	63.62	54.86
BERT [10]	47.70	77.42	76.35	68.80
VL-BERT [42]	49.31	78.13	80.51	71.91
ViLBERT [30]	49.18	78.24	80.51	71.91
VisualBERT [27]	49.20	78.41	81.23	72.39
CLIP [38]	49.39	78.35	81.75	72.50
UNITER [7]	49.87	78.54	82.71	73.58
CAPTURE [38]	50.30	78.69	83.06	74.14
SCALE (Ours)	**51.47**	**79.58**	**84.23**	**75.81**
Image_{based}	22.67	30.14	67.49	59.64
BERT [10]	55.82	82.11	87.30	71.75
CLIP [38]	57.73	82.60	90.49	76.48
SCALE (Ours)	**59.81**	**83.68**	**92.01**	**78.34**

while the remaining transformers are randomly initialized. Both the single-modality encoders and the multi-modal fusion encoders consist of 6 transformer layers each, adding up to a total of 12 transformer layers. The hidden state size of each modality transformer is 768 and the maximum sequence length for the captions and tables are set to 36 and 64, respectively. Using the same setting as in [30] [5], we utilize Faster R-CNN [40] with a backbone ResNet101 [15] pre-trained on the Visual Genome dataset [23] to extract region features of selected 10 to 36 bounding boxes with high-class detection probability for each image. We train **SCALE** with a total batch size of 64 for 5 epochs using the Adam optimizer [21] with a warm-up learning rate of $1e$-4. Additional implementation details of our model are provided in Section G of the supplementary material.

Baselines. We compare **SCALE** to the following eight alternative pre-training methods that utilize image and text modalities as well as combinations of both: Bert [10] (Text_{based}), Image_{based}, ViLBERT [30], CLIP [38], VL-BERT [42], VisualBERT [27], UNITER [7] and CAPTURE [56]. Image_{based} and BERT [10] are 12-layer transformers based on the MLM (Mask Language Modeling) or MRP (Mask Region Prediction) task using image or text modality, providing single-modal baselines for the product retrieval, classification, and clustering tasks. To ensure a fair comparison, the same hidden size of 768 is chosen for all baselines.

Evaluation. We consider the following four downstream tasks to evaluate the learned representations: 1) Multi-modal retrieval: This task aims to find the most relevant target products using combinations of two or more modalities. A pair is considered a match if both belong to the same category; 2) Fine-grained multi-modal retrieval: Retrieval on an instance level, where only samples of the same product (i.e. color, model, shape, and style) are considered

[5] https://github.com/airsplay/py-bottom-up-attention

a match [6]; 3) Multi-modal classification: Product category classification given the multi-modal features extracted from the joint co-transformer of **SCALE** using a linear classifier; and 4) Multi-modal clustering: Product category clustering using k-Means clustering and the same features as in the classification setting. For product retrieval, we adopt the widely used metrics mean Average Precision (mAP) and Precision (Prec) [14, 31, 49] to evaluate the retrieval accuracy on the two retrieval tasks. For product classification and clustering, all methods are evaluated using the Classification Precision (Classification accuracy), Normalized Mutual Information (NMI) [52] and Purity metrics. In all experiments, models are trained on the training split. The pre-trained model is then applied to extract the modality features of the gallery and test splits for the product retrieval and clustering tasks. For the classification task, we finetune the pre-trained model on the classification subset containing 1,805 categories/classes and utilize the finetuned model to extract the features of the classification test set.

5.1. Modality Diversity

To examine the performance of our proposed **SCALE** framework and to verify the benefits of diverse modalities and dataset scale, we train **SCALE** with an increasing number of modalities and observe the variations in classification and multi-modal retrieval performance both for the whole **M5**Product dataset and the subset. More specifically, fused features are extracted from the joint co-transformer (**JCT**) of our **SCALE** after finetuning for the classification task and after pre-training and finetuning for the (coarse) multi-modal retrieval task. Results in Table 3 show that performance increases across all settings as modalities are added, illustrating the benefit of complementary modality information to learning multi-modality feature representations. It can also be observed that modality gains are larger on the whole dataset, supporting *Interesting Observation 1*.

We further provide results for an extensive set of modality combinations to verify **SCALE**s effectiveness in leveraging the diverse modalities of our **M5**Product dataset. Table 4 provides results for the coarse- and fine-grained multi-modal retrieval tasks as well as the classification task after finetuning the model. As in the previous experiment, noticeable improvements can be observed as additional modalities are added. In particular, the addition of the text modality leads to high modality gains, verifying the benefits of including more diverse modalities that can capture different views of the same product. Interestingly, performance on the coarse-grained retrieval task is significantly worse than on the fine-grained retrieval task in most cases, indicating the complexity of the **M5**Product dataset and the diversity of the products in each category.

[6] A more thorough definition of the term *same products* and how instance-level labels are obtained is provided in the supplementary.

Table 6. Ablation study of the **SIMCL** module.

#	IMCL	PRE	Accuracy	mAP@1,5,10	Prec@1,5,10
1			83.77	68.45 / 70.92 / 69.30	67.56 / 46.37 / 34.12
2	✓		84.44	69.14 / 71.96 / 70.13	69.14 / 47.15 / 34.84
3		✓	84.09	69.31 / 71.59 / 69.85	69.31 / 46.72 / 34.42
4	✓	✓	**85.50**	**70.62 / 73.02 / 71.50**	**70.62 / 48.20 / 35.35**

Table 7. Analysis of different masked tasks (token mask (MLM) and entity mask (MEM)) for the table modality.

Tasks	Accuracy	mAP@1,5,10	Prec@1,5,10
MLM	84.05	68.34 / 71.19 / 69.43	68.34 / 47.02 / 34.43
MEM	**85.50**	**70.62 / 73.02 / 71.50**	**70.62 / 48.20 / 35.35**

Table 8. Analysis of treating text and table modalities separately (T/Tab) or stacked together (T+Tab).

Formats	Accuracy	mAP@1,5,10	Prec@1,5,10
T+Tab	84.61	70.15 / 72.19 / 70.49	69.15 / 47.40 / 34.40
T/Tab	**85.50**	**70.62 / 73.02 / 71.50**	**70.62 / 48.20 / 35.35**

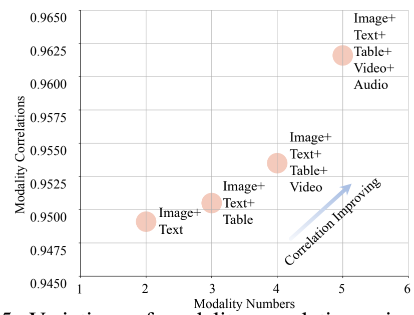
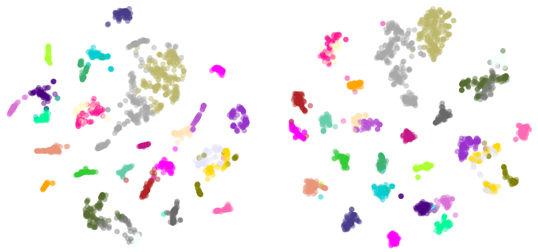


Figure 5. Variations of modality correlation gains with the number of modalities.

Semantic Alignment. To additionally demonstrate the importance of modality diversity, we compute the modality correlation, the average cosine similarity between image and text features as obtained by the **JCT**, for an increasing number of modalities. Figure 5 illustrates that the semantic alignment capability of the pre-training model increases as the number of modalities grows.

5.2. Multi-modal Downstream Tasks

We evaluate **SCALE** on the **M5**Product dataset for the product retrieval, classification, and clustering tasks and compare results to several benchmark approaches in Table 5. For the Image$_{based}$ and BERT [10] models, which only utilize the image and text features, respectively, the extracted features are fed directly into the classification model. For our **SCALE** approach, we utilize the fused modality features generated by the joint co-transformer, pre-trained on both image and text modalities. Only utilizing the image and text modalities allows us to facilitate a fair comparison to the recent state-of-the-art approaches ViLBERT [30], CLIP [38], VL-BERT [42], VisualBERT [27], UNITER [7] and CAPTURE [56]. Comparing our **SCALE** framework to the unimodal models, Image$_{based}$ and Bert [10], we observe that exploiting multi-modal data significantly im-



(a) VILBERT (b) SCALE

Figure 6. Visualize the embeddings generated by **SCALE** and VILBERT via t-SNE. Points belonging to the same category are of the same color. Best viewed in color.

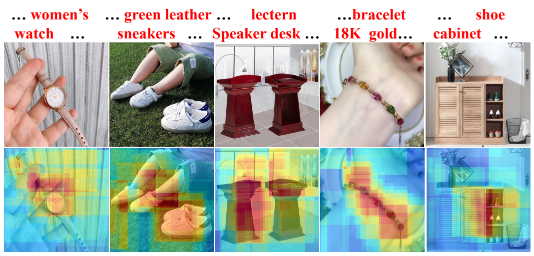


Figure 7. Attention attribution over proposals learned by our **SCALE**.

proves the performance across all tasks. We further observe that **SCALE**, by leveraging **SIMCL**, can efficiently fuse the modalities and outperform all the baseline approaches (*Interesting Observation 2*).

5.3. Ablation Studys and Visualization

To explore how **SIMCL** influences **IMCL** and the **Pre**text tasks, we conduct several ablation studies. Table 6 illustrates that improvements of approximately 2% are obtained in the classification task and more than 2% for the coarse-grained retrieval task when including both, highlighting the importance of both the **Pre**text tasks and the effective modality fusion of **SIMCL**. We further analyze the effect of the MEM pretext task for the table modality and show the benefit of masking out complete entities over masking out individual tokens (MLM) in Table 7. This benefit can be attributed to the fact that MEM ensures that **SCALE** learns representations that encode the semantic information of complete entities. Finally, we evaluate the performance of modelling the text and the table modalities using individual modality encoders and compare **SCALE**s retrieval performance to a baseline where text and table information is concatenated and fed to a single transformer, resembling the process of BERT [10]. By modelling both modalities individually, results in Table 8 illustrate that more information can be preserved and we hypothesize that using a single transformer leads to a loss in table modality information for the benefit of the more expressive text modality.

Figure 6 shows t-SNE visualizations of the extracted features for the **JCT** module of our **SCALE** model and the alternative approach ViLBert [30] on the **M5**Product dataset. **SCALE** not only better distinguishes different classes but also improves class compactness compared to the ViLBert model. Further, the attention attribution for different modalities are shown in Figure 7 and verify that the visual features generated by **SCALE** are object-oriented and semantically interpretable.

6. Limitations and future work

The experimental evaluation showed that **SCALE** is able to learn efficient representations from a large number of modalities for retrieval, classification, and clustering. However, more evaluation of the generative capabilities of the models representations is lacking and tasks such as image and caption generation could be promising directions to explore. We further provide some of **SCALE**s failure cases in supplementary Section J.

Potential negative societal impact. As a result of the strict ethical considerations used in the data collection process, where among others personally identifiable information has been removed, **M5**Product does not pose any ethical risks.

7. Conclusion and Discussion

To facilitate multi-modal pre-training, we present the **M5**Product dataset, which is the largest available multi-modal E-commerce product dataset, consisting of five core modalities (image, text, table, video, and audio). To further promote multi-modal research in retail and increase seller and buyer engagement, we also propose the novel **SCALE** multi-modal pre-training framework. By utilizing Self-harmonized Inter-Modality Contrastive Learning (**SIMCL**), **SCALE** is able to model and exploit modality relationships effectively and outperforms previous approaches on the **M5**Product multi-modal retrieval, classification, and clustering tasks. We believe that both the dataset and the proposed framework work will inspire research on scaling multi-modal pre-training beyond the commonly used image and text modalities.

8. Acknowledgement

This work was supported in part by Shenzhen Fundamental Research Program (Project No. RCYX20200714114642083, No. JCYJ20190807154211365), National Key RD Program of China (2021ZD0112100), and CAAI-Huawei MindSpore Open Fund . We thank MindSpore for the partial support of this work, which is a new deep learning computing framwork[7], and supported by Guangdong Provincial Key Laboratory of Fire Science and Intelligent Emergency Technology, Guangzhou 510006, China.

[7] https://www.mindspore.cn/

References

[1] Open images dataset. https://storage.googleapis.com/openimages/web/index.html/, 2018. 2, 14

[2] Huda AlAmri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, and Devi Parikh. Audio visual scene-aware dialog. In *CVPR*, pages 7558–7567, 2019. 2, 14

[3] 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 *CVPR*, pages 6077–6086, 2018. 4

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, pages 2425–2433, 2015. 2, 14

[5] 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 *CVPR*, pages 3558–3568, 2021. 2, 3, 14

[6] Delong Chen, Fan Liu, Xiaoyu Du, Ruizhuo Gao, and Feng Xu. Mep-3m: A large-scale multi-modal e-commerce products dataset. 2021. 3, 14

[7] 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 Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer, 2020. 2, 3, 6, 7

[8] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: a real-world web image database from national university of singapore. In *CIVR*, 2009. 2, 3, 14

[9] Charles Corbiere, Hedi Ben-Younes, Alexandre Ramé, and Charles Ollion. Leveraging weakly annotated data for fashion image retrieval and label prediction. In *ICCV Workshops*, pages 2268–2274, 2017. 2, 3, 14

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5, 6, 7, 8, 12

[11] Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*, 2020. 3

[12] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, volume 12349, pages 214–229. Springer, 2020. 3

[13] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *SIGIR*, pages 2251–2260, 2020. 3

[14] Yunchao Gong, Svetlana Lazebnik, Albert Gordo, and Florent Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12):2916–2929, 2013. 7

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6, 12

[16] Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *EMNLP*, pages 861–877, 2020. 1

[17] Yuting Hu, Liang Zheng, Yi Yang, and Yongfeng Huang. Twitter100k: A real-world dataset for weakly supervised cross-media retrieval. *IEEE Trans. Multim.*, 20(4):927–938, 2018. 2, 14

[18] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *arXiv preprint arXiv:2106.04538*, 2021. 1

[19] Keith Ito and Linda Johnson. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017. 2, 14

[20] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: toward spatio-temporal reasoning in visual question answering. In *CVPR*, pages 1359–1367, 2017. 2, 14

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6, 14

[22] Josip Krapac, Moray Allan, Jakob J. Verbeek, and Frédéric Jurie. Improving web image search results using query-relative classifiers. In *CVPR*, pages 1094–1101, 2010. 2, 14

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017. 2, 6, 14

[24] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. TVQA: localized, compositional video question answering. In *EMNLP*, pages 1369–1379, 2018. 2, 14

[25] Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. In *ISCA*, pages 3459–3463, 2018. 2, 3, 14

[26] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. HERO: hierarchical encoder for video+language omni-representation pre-training. In *EMNLP*, pages 2046–2065, 2020. 3

[27] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019. 2, 3, 6, 7

[28] Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, Jie Zhang, Jianwei Zhang, Xu Zou, Zhikang Li, Xiaodong Deng, Jie Liu, Jinbao Xue, Huiling Zhou, Jianxin

Ma, Jin Yu, Yong Li, Wei Lin, Jingren Zhou, Jie Tang, and Hongxia Yang. M6: A chinese multimodal pretrainer. *CoRR*, abs/2103.00823, 2021. 3

[29] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014. 2, 14

[30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NIPS*, pages 13–23, 2019. 2, 3, 6, 7, 8

[31] Xiaoqiang Lu, Xiangtao Zheng, and Xuelong Li. Latent semantic minimal hashing for image retrieval. *IEEE Trans. Image Process.*, 26(1):355–368, 2017. 7

[32] 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 *ICCV*, pages 2630–2640, 2019. 2, 3, 14

[33] Ksr Murty and B. Yegnanarayana. Combining evidence from residual phase and mfcc features for speaker recognition. *IEEE Signal Processing Letters*, 13(1):52–55, 2005. 4, 13

[34] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, pages 1143–1151, 2011. 2, 14

[35] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS Workshop*, 2017. 14

[36] Yuxin Peng, Xin Huang, and Yunzhen Zhao. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *IEEE Trans. Circuits Syst. Video Technol.*, 28(9):2372–2385, 2018. 3, 14

[37] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020. 3

[38] 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 *ICML*, pages 8748–8763, 2021. 2, 3, 6, 7

[39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 2021. 3

[40] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017. 6

[41] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018. 1, 2, 3, 14

[42] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: pre-training of generic visual-linguistic representations. In *ICLR*, 2020. 2, 3, 6, 7

[43] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, pages 6418–6428, 2019. 2, 14

[44] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, pages 7463–7472, 2019. 3

[45] Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In *EMNLP*, pages 5099–5110, 2019. 2, 3

[46] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, pages 4631–4640, 2016. 2, 14

[47] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, pages 4580–4590, 2019. 2, 14

[48] Xiu-Shen Wei, Quan Cui, Lei Yang, Peng Wang, and Lingqiao Liu. Rpc: A large-scale retail product checkout dataset. *arXiv preprint arXiv:1901.07249*, 2019. 2, 3, 14

[49] Yair Weiss, Antonio Torralba, and Robert Fergus. Spectral hashing. In *NIPS*, pages 1753–1760, 2008. 7

[50] Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier R. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *NIPS*, pages 2035–2043, 2009. 12

[51] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016. 2, 14

[52] Chengfu Yang and Zhang Yi. Document clustering using locality preserving indexing and support vector machines. *Soft Comput.*, 12(7):677–683, 2008. 7

[53] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78, 2014. 2, 14

[54] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *ACL*, pages 2236–2246, 2018. 3, 14

[55] Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *ICCV*, 2021. 2, 3, 14

[56] Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. *CoRR*, abs/2107.14572, 2021. 2, 6, 7

[57] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, pages 7590–7598, 2018. 2, 14

[58] Tao Zhou, Mingxia Liu, Huazhu Fu, Jun Wang, Jianbing Shen, Ling Shao, and Dinggang Shen. Deep multi-modal latent representation learning for automated dementia diagnosis. In *MICCAI*, pages 629–638, 2019. 1

[59] Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. Kaleido-bert: Vision-language pre-training on fashion domain. In *CVPR*, pages 12647–12657, 2021. 3