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

A. Dataset License

Our **M5**Product dataset is released under CC BY-NC-SA 4.0 license and can freely be used for non-commercial purposes. More detailed information can be found at https://xiaodongsuper.github.io/M5Product_dataset/terms_of_use.html, which also provides dataset details and usage guidance.

B. Annotation Collection

We resort to crowd-sourcing to obtain human annotations for the product retrieval task. Specifically, we present human annotators with a matching task, where annotators are asked to select the matching image-text pairs for a given query image-text pair. In our crowdsourcing system, each matching task is presented to five different human annotators and a typical example of our interface is shown in Figure 1. The left part of the interface shows the current query data (image and text), while the right side depicts an example from the candidate list. The annotators are then asked to choose from two options: *mismatched* and *uncertain*. The default labelling option is *matched*. The interface also displays the number of examples that have been reviewed and the total amount of examples to review. Each annotation task, can be considered as a binary classification task for the human worker, where he/she has to decide if the pair is a match or not. For each estimated task, the annotators receive a payment of 3 cents RMB.

C. Annotation

The retrieval task annotation for any query sample consists of all the matched instances in the gallery split. To construct a reliable gallery set, we first use a ResNet50 [9] and Bert-Base [8] to extract features and construct the query candidate pool from all the data that is not contained in the training subset. Specifically, we sample an instance from a category that contains more than 2,000 instances and extract the image and text features. We then concatenate the features and compute the cosine similarity to all other instances of the dataset to produce a pre-ranked candidate list in order to minimize the labelling cost. The final size of the candidate shortlist for each query is 500, which is about 0.01% of the whole gallery split. During the crowd-sourced annotation process, human workers review both images and captions in the candidate list to select which samples are matched with the query instance.

Annotation Rules. It is quite challenging to define whether two images contain the same product when critical aspects are not given in their captions and images. In our annotations, we use product images and their captions as the primary materials for gallery construction. Hence, we define several rules to determine the "same product" condition and provide them as instructions to the annotators. Images contain the same product, if:

1. The two images are in different conditions (e.g., backgrounds, angles, etc), but the products in both images are the same.
2. They should have the same color/model/shape/style, or other features that can be distinguished by humans.
3. The caption has the same product name but the product description differs.
4. They share more than one characteristic such as appearances, materials, colors and so on.

To ensure labeling consistency, each annotation pair is labeled by five human workers in the crowd-sourced platform. In the process, we first make a small dataset from our query list as a Gold Problem to evaluate the annotation capability of each human worker. Based on the labeled results ("Matched" or "Not Matched") from human workers and their annotation capability, we utilize the weighted

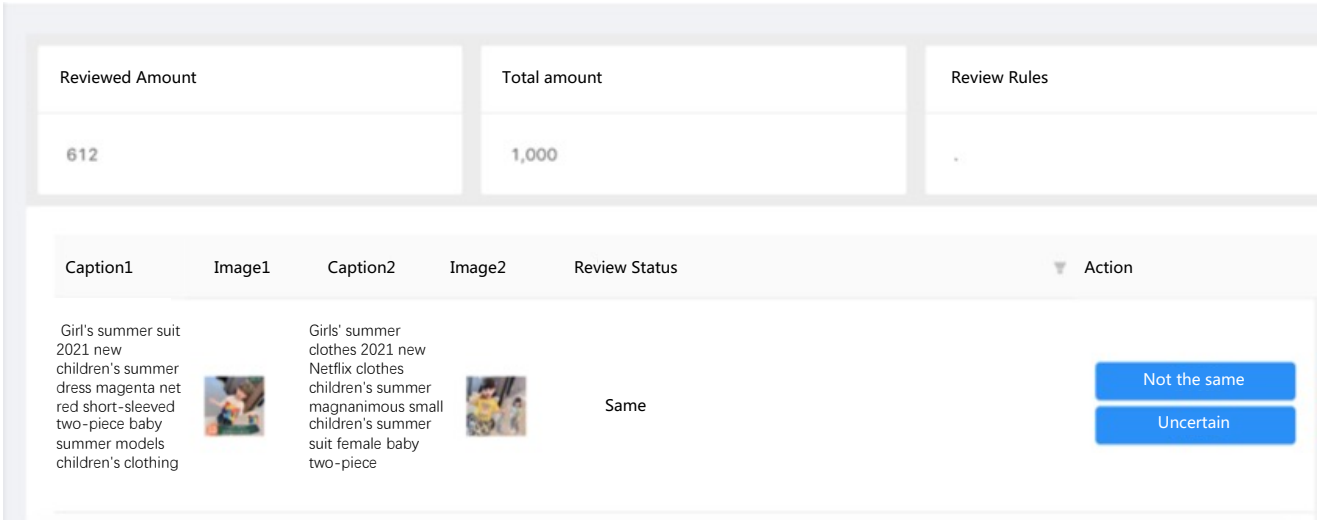

Figure 1. UI for huamn annotation on product retrieval task.

GLAD [29] inference algorithm to determine the final accepted labels.

D. Dataset Split

The **M**5Product dataset is split into several parts to ensure consistent training and evaluation of the models on the various tasks. The *training* set contains 4,423,160 samples from 3,593 classes.

Retrieval To evaluate models on the retrieval tasks, the remaining data is split into *gallery-c* and *query-c* sets, which are used for the coarse-grained retrieval task, and *gallery-fg* and *query-fg* sets, which are used for the fine-grained retrieval task. The difference between the two retrieval tasks lies in the granularity of their annotation. In the fine-grained task, only identical products are considered a match (for example, all IPHONE 11 Black), while in the coarse-grained task, category labels are being used to group products from each category (for example, all phones are considered a match).

To construct the fine-grained sets, we extracted all cosmetics categories and, using the abovementioned annotation procedure, finally obtained 1,991 *query-fg* samples and 117,858 *gallery-fg* samples. The *query-c* and *gallery-c* sets contain 24,410 and 1,197,905 samples, respectively. Among the samples in the *gallery-c* set, 249,614 samples are matched with samples in the *query-c* set, while 948,291 samples do not match. These unmatched samples are added to the *gallery-c* set to increase the difficulty of the retrieval task. We further report the finetuned retrieval performance in the paper, which corresponds to retrieval performance after finetuning the model using the classification training set (see next paragaraph).

Finetuning, Classification and Clustering For the classification and clustering tasks, we sampled 1,805 categories from the whole dataset and obtained 18,526 train samples and 4,632 test samples. We first finetune our **SCALE** using the classification training set and then extract features from the finetuned model to perform the classification and clustering tasks.

E. Data Format

The dataset consists of 6,313,067 products uploaded by 1,000,517 merchants, where merchant information has been removed to ensure anonymity. In the following, we outline the different modalities:

Image data Each product has at least five product images, where the first image is the main image that gives the detailed overview of the product, while the rest depict its functionalities or characteristics. We pick all the main images to construct the dataset.

Caption/text data are provided by the 1,000,517 merchants. Note that the text description does not always match well with the other modalities.

Video data are used to showcase the products' usage and characteristics to customers. In our dataset, these videos are recorded at a speed of 24 frames per second (FPS). To reduce the amount of redundant information that is contained in adjacent frames and the dataset as a whole, we only select one frame per second.

Audio data are extracted from the video data. We extract the corresponding audio information of all sampled video frames. Then the audio frames are transformed into spectrograms using Mel-Frequency Cepstral Coefficients (MFCC) [20]. We set the frame size and hop size as 1,024 and 256, respectively.

Tabular data are a special kind of database that records some additional product characteristics such as appearance, purpose and producer. The tabular data is indexed by the product ID and collected from the whole product database. There are 5,679 different types of property information and 24,398,673 unique values.

Table 1. More results on SOP datasets (blue) and Caltech101 (red).

Method	mAP@1	mAP@5	mAP@10	Accuracy
ViLBERT	36.87	39.47	38.42	97.81
UNITER	37.31	40.91	39.38	**98.49**
CAPTURE	38.24	41.37	40.67	98.14
SCALE (Ours)	**39.13**	**42.67**	**41.92**	**98.49**

Table 2. Comparisons on the MSVD video dataset.

Method	Text to Video		
	R@1	R@5	R@10
HERO	16.8	43.4	57.7
FROZEN	33.7	64.7	76.3
CLIP	37.0	64.1	73.8
SCALE (Ours)	**40.3**	**69.9**	**78.1**

F. Unimodal and unpair analysis

1) **Unimodal analysis**: Figure 2 gives the video, text and merchant distributions. From the figure, we can find that the video duration, the text length and the merchant number range from 1 to 60 seconds, 20 to 40 words and 1 to 10 product numbers, respectively. This variation further illustrates the real-world nature of our dataset. 2) **Unpair analysis**: In the data collection process, there are 82,577 invalid URLs for the image modalities (1.3% of the products), while the number of samples that contain both the Image and Text modalities is 6,230,490. Further taking into account the table modality, the number of complete samples drops by 1.4% to 6,225,598 samples that have all three modalities. Overall, the dataset contains 5,050,078 samples that contain all five modalities. This means that about 20% of the samples are incomplete. This is mostly due to merchants being biased towards specific modalities, which is a common scenario in the real world.

G. Implementation Details

Our models are implemented in Pytorch [22]. To speed up training, we use Nvidia Apex[1] for mixed precision training. All models are trained on 4 Nvidia 3090 and 2080ti GPUs on our workstations. We use Adam [13] to optimize the parameters of our model, with an initial learning rate of 1**e**-4, and use a linear learning rate decay schedule with a temperature parameter of 0.1.

[1] https://github.com/NVIDIA/apex

H. More performance verification

Generalization verification. We compare our **SCALE** to several alternatives on the Caltech101 and Stanford Online Products (SOP) datasets for image classification and retrieval tasks, respectively. Meanwhile, we also provide a comparison on the video dataset MSVD for further verifying the effectiveness of our model **SCALE**. The corresponding results are shown in Table 1 and 2. The results verify that our model, pretrained using the M5Product dataset, can achieve several tasks in the general domain besides the fashion/e-commercial domain.

SOTA image-text approaches. Due to the specific network design for the cross-modal interaction, neither bimodal nor trimodal based methods can be directly applied to our dataset. To ensure fair evaluation, we therefore compare performances for the image and text modalities. Here, we try to adjust the architecture of CLIP for three modalities. The mAP results are shown in Table 5.

Performance comparisons with more weights. Due to the specific design of different SOTA models for different modality inputs, it is infeasible to adjust their model to fit our M5Product. We select to assign more weights to the text modality for the baseline models and show the results in Table 6.

I. Dataset Comparison

A comprehensive comparison between our **M5**Product dataset and other widely used multi-modal pre-training datasets is shown in Table 3. From the table, we can observe that our **M5**Product not only has more diverse modalities but also contains a large amount of data samples from an abundant amount of categories.

J. Missing data verification

Results in Table 4 show the superiority of our methods over the standard approach of ignoring incomplete samples. We compare two variants of our **SCALE** framework: 1) **SCALE** (full-modality) and our proposed **SCALE**. The only difference between the two methods is the input. The input of the former only includes complete samples (all modalities present), while the input of the latter includes the incomplete modality samples. The verification is performed on the subset dataset as mentioned in the main article.

K. Failure Analysis

Several product retrieval examples are shown in Figures 3, 4, and 5. The first column represents the image and text modality of the query sample, while the eight images to its right belong to the matched results from the gallery set. In the matched results, the samples boxed in blue are the correctly matched samples, while the samples boxed in red

Table 3. Comparisons with other widely used multi-modal datasets. "-" means not mentioned. Our **M**5Product is one of the largest multi-modal datasets compared with existing datasets.

Dataset	Samples	Categories	Instances	Modalities	Modal type	Product
LJ Speech [11]	13,100	-	-	2	audio/text	no
SQuAD [17]	37,111	-	-	2	audio/text	no
TVQA [16]	21,793	-	-	2	video/text	no
MovieQA [26]	408	-	-	2	video/text	no
TGIF-QA [12]	56,720	-	-	2	video/text	no
AVSD [2]	11,816	-	-	2	video/text	no
Youcook2 [34]	14,000	89	-	2	video/text	no
VATEX [27]	35,000	-	-	2	video/text	no
MSRVTT [30]	100,000	20	-	2	video/text	no
HowTo100M [19]	1,220,000	12	-	2	video/text	no
Conceptual Caption 3M [24]	3,300,000	-	-	2	image/text	no
SBU [21]	890,000	-	-	2	image/text	no
Visual Genome [15]	108,000	-	-	2	image/text	no
COCO [18]	123,287	-	-	2	image/text	no
Flickr30K [31]	31,000	-	-	2	image/text	no
NLVR2 [25]	107,292	-	-	2	image/text	no
VQA2.0 [3]	204,721	-	-	2	image/text	no
RPC checkout [28]	30,000	200	367,935	2	image/text	no
Twitter100k [10]	100,000	-	-	2	image/text	no
INRIA-Websearch [14]	71,478	353	-	2	image/text	no
NUS-WIDE [6]	269,648	81	-	2	image/text	no
Open Image [1]	1,670,000	-	-	2	image/text	no
Conceptual 12M [4]	12,423,374	-	-	2	image/text	no
CMU-MOSEI [32]	23,500	2	-	3	text/video/audio	no
XMedia [23]	12,000	20	-	5	image/text/video/audio/3D	no
Dress Retrieval [7]	20,200	50	∼20,200	2	image/text	yes
MEP-3M [5]	3,012,959	599	-	2	image/text	yes
Product1M [33]	1,182,083	458	92,200	2	image/text	yes
M5Product	**6,313,067**	**6,232**	-	**5**	**image/text/video/audio/table**	**yes**

Table 4. The retrieval performance with missing modalities.

Modal	Accuracy	mAP@1	mAP@5	mAP@10	Prec@1	Prec@5	Prec@10
SCALE (full-modality)	84.06	57.97/ 69.12	62.54/ 71.93	60.48 / 69.92	57.97 / 69.12	38.02 / 47.63	28.88 / 34.70
SCALE	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

Table 5. The results using more modalities with SOTA method.

Method	mAP@1	mAP@5	mAP@10
CLIP-ITTab	67.23	69.12	68.72
SCALE (Ours)	**67.97**	**70.34**	**69.38**

Table 6. Performances with more weights for the text modality.

Text:Image (3:2)	mAP@1	mAP@5	mAP@10
ViLBERT	49.93	54.55	52.45
UNITER	49.28	54.14	53.48
CAPTURE	50.57	55.27	53.18
SCALE (Ours)	**51.47**	**56.16**	**54.41**

are mismatched. These retrieval results illustrate that the learned embeddings are discriminative. However, in a few cases, the recalled samples are not matched due to the limited number of category samples in the gallery set or similar descriptions in the text data.

L. More Visualization

Additional attention visualization results are provided in Figures 6 and 7. Similar to the illustrations in the main

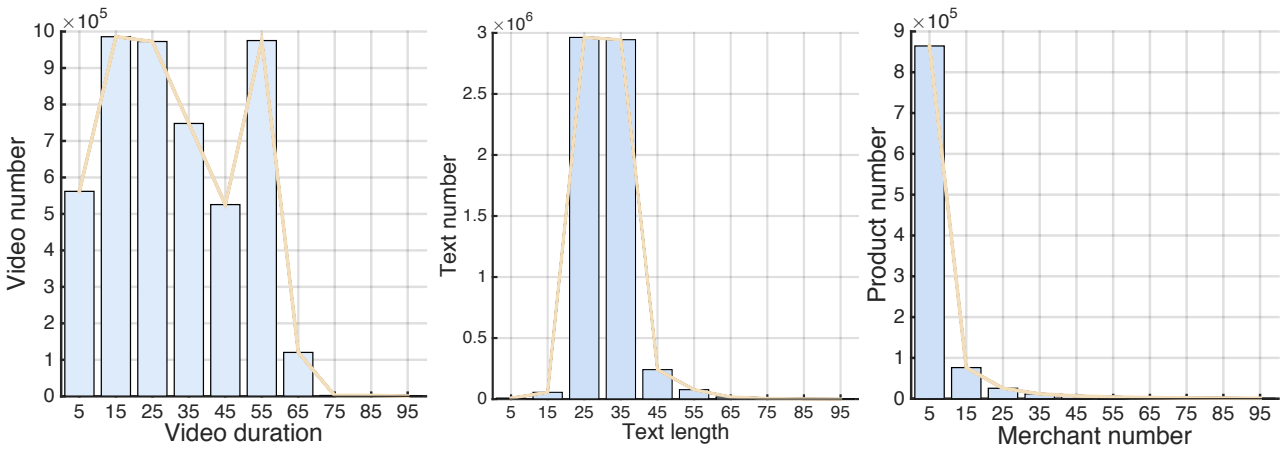

Figure 2. The distributions of video, text and merchant on our **M5**Product.

Car silk circle foot mats Changan Ease cs35 new cx20 Yuexiang V5 to Shang XT Ben Ben mini main driver single piece

Zhang Xiaoquan kitchen knife home stainless steel slicing knife chef special kitchen knife cutting vegetables and meat free grinding kitchen knives

Figure 3. Successful retrieval results 1 by our **SCALE**.

Figure 4. Failure retrieval results 2 by our **SCALE**.

paper, these visualizations show that **SCALE** can learn the detailed semantics in the images and the text.

References

[1] Open images dataset. https://storage.googleapis.com/openimages/web/index.html/, 2018. 4

[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. 4

[3] 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. 4

[4] 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. 4

[5] Delong Chen, Fan Liu, Xiaoyu Du, Ruizhuo Gao, and Feng Xu. Mep-3m: A large-scale multi-modal e-commerce products dataset. 2021. 4

[6] 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. 4

[7] 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. 4

[8] 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. 1

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*,

TOMY Domeka alloy car model male toys TOMICA Ferrari 246GT / F40 / 512BB / 365GT

Yun Yun inter Anji white tea 2020 new tea Ming Qian first pick authentic master spring tea premium rare bulk 100g

Figure 5. Failure retrieval results 3 by our **SCALE**.

pages 770–778, 2016. 1

[10] 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. 4

[11] Keith Ito and Linda Johnson. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017. 4

[12] 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. 4

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 3

[14] 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. 4

[15] 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. 4

[16] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. TVQA: localized, compositional video question answering. In *EMNLP*, pages 1369–1379, 2018. 4

[17] 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. 4

[18] 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. 4

[19] 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. 4

[20] Ksr Murty and B. Yegnanarayana. Combining evidence from residual phase and mfcc features for speaker recognition.

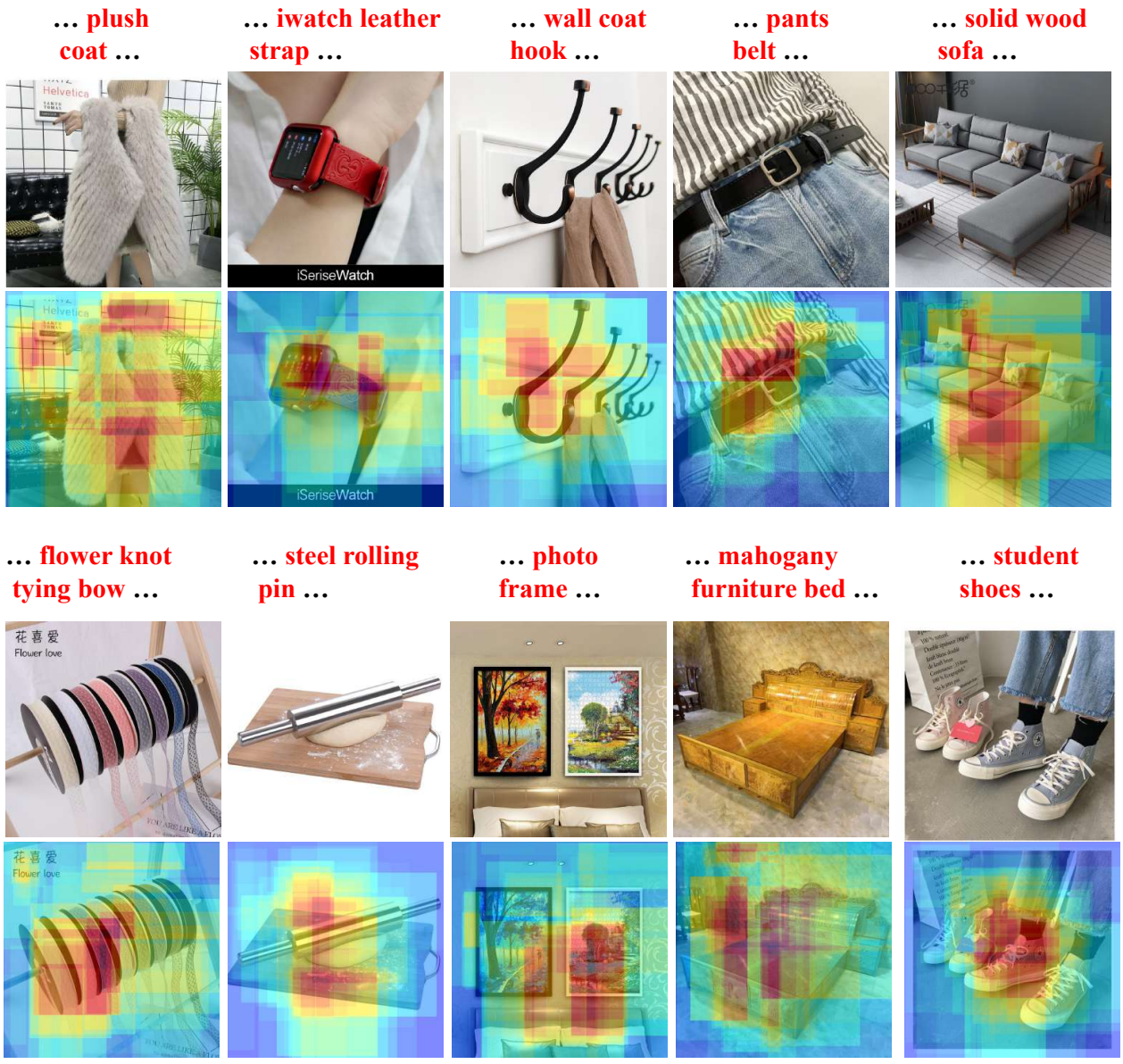

Figure 6. More attention visualization 1 by our **SCALE**.

IEEE Signal Processing Letters, 13(1):52–55, 2005. 2

[21] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, pages 1143–1151, 2011. 4

[22] 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. 3

[23] 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. 4

[24] 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. 4

[25] 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. 4

[26] 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. 4

[27] 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. 4

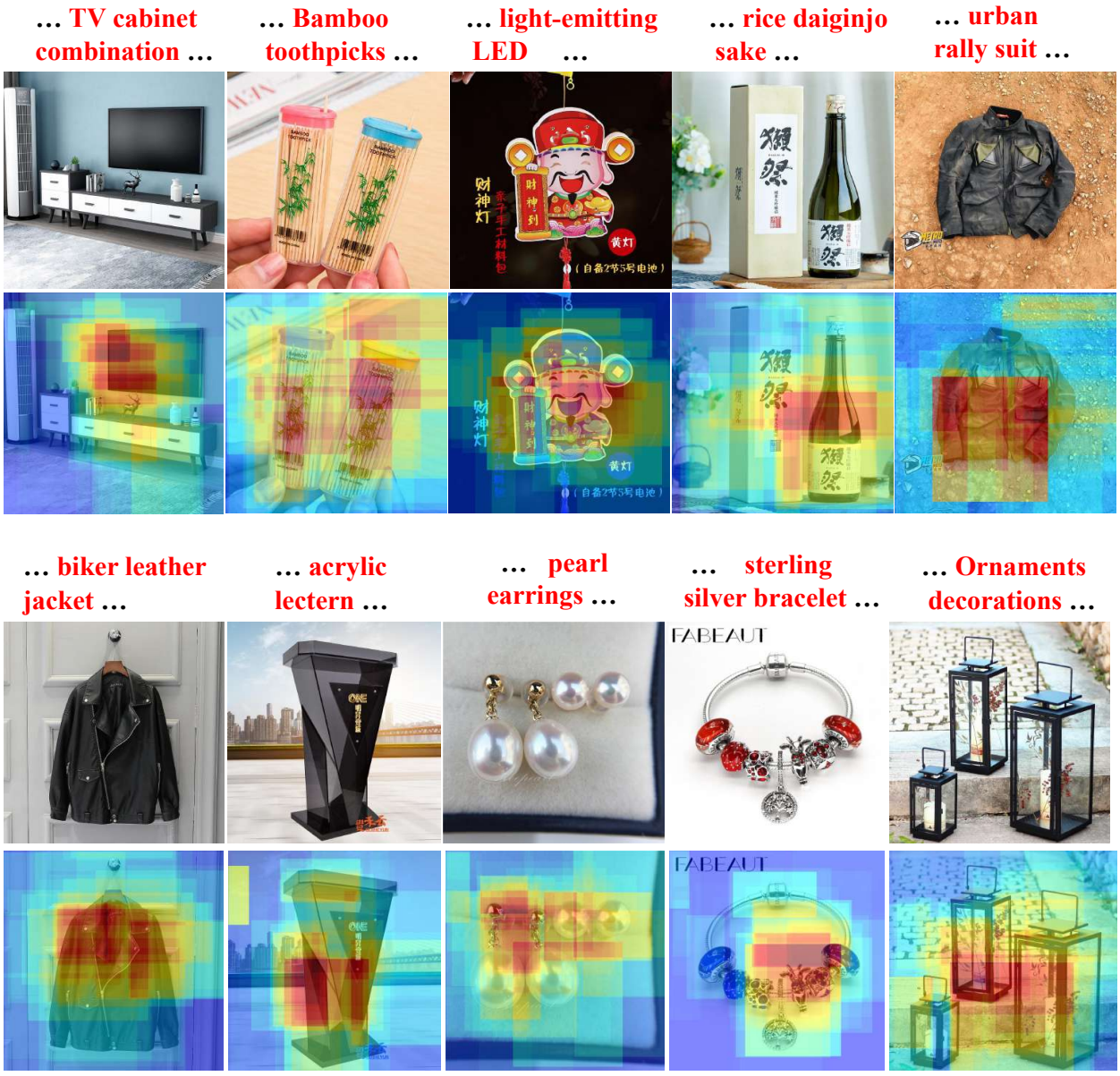

Figure 7. More attention visualization 2 by our **SCALE**.

[28] 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. 4

[29] 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. 2

[30] 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. 4

[31] 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. 4

[32] 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. 4

[33] 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. 4

[34] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, pages 7590–7598, 2018. 4