KAFA: Rethinking Image Ad Understanding with Knowledge-Augmented Feature Adaptation of Vision-Language Models

Zhiwei Jia* UC San Diego
Pradyumna Narayana Google
Arjun R. Akula Google
Garima Pruthi Google
Hao Su UC San Diego
Sugato Basu Google
Varun Jampani Google

Abstract

Image ad understanding is a crucial task with wide real-world applications. Although highly challenging with the involvement of diverse atypical scenes, real-world entities, and reasoning over scene-texts, how to interpret image ads is relatively under-explored, especially in the era of foundational vision-language models (VLMs) featuring impressive generalizability and adaptability. In this paper, we perform the first empirical study of image ad understanding through the lens of pre-trained VLMs. We benchmark and reveal practical challenges in adapting these VLMs to image ad understanding. We propose a simple feature adaptation strategy to effectively fuse multimodal information for image ads and further empower it with knowledge of real-world entities. We hope our study draws more attention to image ad understanding which is broadly relevant to the advertising industry.

1 Introduction

As advertisements play an integral role in human society, image ad understanding has many real-world applications such as ad targeting (Hussain et al., 2017), visual metaphor understanding (Abokhoza et al., 2019) and creative ad generation (Chilton et al., 2019; Akula et al., 2022). It is also highly challenging due to several reasons, as exemplified in Fig. 2. *First*, image ads consist of diverse visual elements including non-photorealistic objects and atypical scenes synthesized creatively that are beyond common academic datasets. *Secondly*, they involve knowledge of a large number of real-world entities such as brands and products where existing work (Su et al., 2018; Li et al., 2022a) struggles to cover. *Lastly*, many adopt visual or multimodal rhetorics requiring reasoning over diverse visual elements including scene-texts, and sometimes even elude humans (Petridis and Chilton, 2019). However, image ad understanding is relatively under-explored in the machine learning community, especially in the presence of recently developed foundational vision-language models (VLMs) pre-trained using a tremendous number of image and text description data.

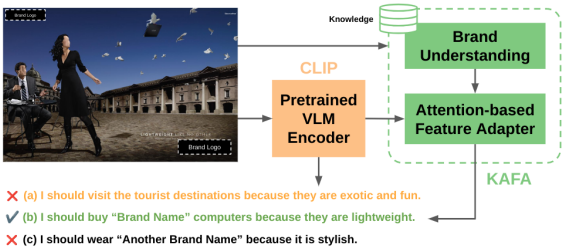

Figure 1: We propose to utilize external knowledge via a brand understanding module and combine features of different modalities via a lightweight attention-based feature adapter to decode the correct messages of image ads. The VLM baseline is confused and gives the wrong one. All brand info is anonymized.

The pre-trained VLMs are shown to have great generalization capability, contain real-world knowledge (implicitly), and can be adapted to a wide range of downstream tasks in a data-efficient way (Radford et al., 2021; Alayrac et al., 2022). It is then natural to utilize VLMs for image ad understanding. In this paper, we perform the first empirical study of adapting VLMs to the task of decoding the overall messages delivered by image ads, which is usually formulated as visual question answering (Hussain et al., 2017). Specifically, we examine three popular pre-trained VLMs that are alignment-based and are publicly available, namely, CLIP (Radford et al., 2021), ALBEF (Li et al., 2021) and LiT (Zhai et al., 2022). We examine zero-shot performance as well as adaptation strategies and reveal the practical challenges of applying VLMs to image ads. We propose a simple feature adaptation strategy that effectively utilizes VLM features. We further propose to incorporate external brand knowledge (real-world entities) that brings a signif-

* Work done in part during an internship at Google. Correspondence to zjia@eng.ucsd.edu.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 5: Industry Track, pages 772–785

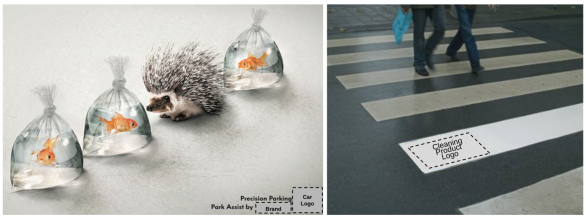

Figure 2: Example image ads with diverse visual elements, atypical scenes and rhetorics to convey their messages creatively. All brand info is anonymized.

icant performance boost.

Our contributions are three-fold. **First**, we empirically find that the sheer scale of data & capacity of the model used in pretraining matters the most for the performance of image ad understanding, partly due to VLM's capability of storing real-world knowledge, which is not captured well by the commonly used metrics for comparing VLMs. **Second**, we reveal the practical challenges of adapting VLMs for image ad understanding (i.e., overfitting to the limited training data & supervision signals and high computation burden of hard negative mining) and propose a simple solution (attention-based feature adaptation) that better leverages VLM features than previous adaptation strategies. **Lastly**, we propose to leverage external knowledge for brand understanding that we have empirically shown to further enhance image ad understanding. Together with the aforementioned adaptation strategy, we call our approach knowledge-augmented feature adaptation (KAFA).

2 Related Work

Image Ad Understanding Learning to automatically interpret image ads was proposed by the Pitt Image Ads Dataset (Hussain et al., 2017), where each ad is annotated by a caption that answers "what should I do according to the ad and why?" Different from traditional image captioning, this task is highly non-trivial as discussed at the beginning of Sec. 1. While prior methods utilize cultural connotations via external symbolic knowledge (Ye and Kovashka, 2018), capture relations between scene-texts and objects by GNNs (Dey et al., 2021), and leverage pre-trained language models to combine multimodel information (Kalra et al., 2020), none have exploited vision-language models (VLMs) and the knowledge of real-world entities (i.e., brands). Besides the wide applications in the ad industry, later work hints that the study of image ads is relevant to much broader research topics (Singh et al., 2019; Akula et al., 2022).

Foundational Alignment-based VLMs A recent surge of collections of tremendous images paired with text descriptions (Schuhmann et al., 2022) enables alignment-based pretraining (i.e., contrastive learning) of foundational VLMs that are efficient zero-shot or low-shot learners for downstream tasks. By learning to embed images and texts into a shared semantic space, they handle domain variations in an open-vocabulary manner (which involves real-world knowledge). Among these are CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), LiT (Zhai et al., 2022) and BASIC (Pham et al., 2021). Another line of work further adopts masked language modeling, image captioning loss, and object-level alignment, e.g., ALBEF (Li et al., 2021), Florence (Yuan et al., 2021), CoCa (Yu et al., 2022) and GLIP (Li et al., 2022b).

Transfer Learning of VLMs Transfer learning of VLMs has become popular with the zero-shot performance of CLIP in image classification tasks. A direct approach is to (partially) fine-tune the VLMs with (optionally) additional neural networks tailored for downstream tasks, e.g., TAP-C (Song et al., 2022), CPT (Yao et al., 2021), KAT (Gui et al., 2021) and VL-Adapter (Sung et al., 2022). Another approach that bypasses the need of tuning the VLMs is prompt learning. For instance, CoOp (Zhou et al., 2022b) and CoCoOp (Zhou et al., 2022a) only tune learnable inputs to the VLMs. The third approach that further reduces memory and computation burden is feature adapters, where VLM features of the inputs are pre-computed before transfer learning. Examples are CLIP-Adapter (Gao et al., 2021), SVL-Adapter (Pantazis et al., 2022) and Attention-Adapter (Zhao et al., 2022).

Knowledge-Augmented Image Understanding Many image understanding tasks require real-world knowledge beyond what can be captured by the input data. For instance, FVQA (Wang et al., 2017) and OK-VQA (Marino et al., 2019) require models to process external fact-based knowledge; TextVQA (Singh et al., 2019) asks to understand named entities in the wild; the Pitt Dataset (Hussain et al., 2017) involves recognition of large quantities of brands. Existing work incorporates external knowledge either explicitly via structured or unstructured knowledge base (Wang et al., 2015; Gardères et al., 2020; Ye and Kovashka, 2018),

or implicitly from knowledge stored in pretrained models (Kalra et al., 2020; Kim et al., 2022), or both (Marino et al., 2021; Gui et al., 2021).

3 What Really Matters for Pre-trained VLMs in Image Ad Understanding?

The first insight of our empirical study is that the sheer size of data and the model used in pretraining is the key factor determining the performance of VLMs for image ad understanding.

To promote reproducibility, we evaluate three alignment-based VLMs (i.e., CLIP, ALBEF and LiT) that are publicly accessible in a zero-shot manner on the Pitt Dataset (Hussain et al., 2017), which formulates ad understanding as image-to-text retrieval. We adopt the official evaluation protocol, which asks the model to select one of the 3 correct messages conveyed by the image ad from a set of 15 candidates (including 12 wrong messages) for each of the 12805 test samples. Specifically, given an alignment-based VLM, let us denote its encoders with normalized outputs as $f_I(\cdot)$ and $f_T(\cdot)$ for image and text branches, respectively. Given an image $\mathbf{x}$ and the ground truth texts $\mathbf{y}$, the VLM retrieves y from candidates $\mathcal{C}(\mathbf{x})$ according to the dot-product score $f_I(\mathbf{x}) \cdot f_T(y)$. We then measure the performance of the VLM with 3 metrics commonly used in the literature: *accuracy* (the percentage of images with any positive text retrieved with rank one), *rank* (how the top retrieved positive text is ranked averaged over all images), and the *mean rank* (the mean rank of the all positive texts averaged over all images).

With the results reported in Tab. 1, we have several findings. *First*, the more data used during the pretraining of a VLM, the better it generalizes to the image ad domain. For a comparison, CLIP has seen 400M image-text pairs, LiT 100M, and ALBEF 14M. *Second*, the larger the capacity of a model, the better it understands image ads. We have evaluated different sizes of the CLIP model beyond the three sizes shown in Tab. 1 and the trend keeps the same. *Third*, commonly used metrics for comparing VLMs, including zero-shot accuracy on the ImageNet (Russakovsky et al., 2015) validation set (for which LiT claims to outperform CLIP) and image-to-text retrieval precision on Flickr30K (Young et al., 2014) (for which ALBEF claims to outperform CLIP), do not reflect the performance of image ad understanding well.

We hypothesize that this is partly because image ad understanding requires knowledge of real-world entities (e.g., brands) which the pre-trained models contain. Similar to the dramatic performance advancement of GPT language models (Brown et al., 2020) driven by the larger scale of training data and the model capacity, more knowledge can be distilled and implicitly stored in the weights of pre-trained VLMs with larger models and more pre-training data. We empirically verify that the VLM's capability of recognizing brands from images is aligned with its performance of decoding the messages from the ads. See results in Tab. 4.

	Acc ↑	Rank ↓	m. Rank ↓
ViLBERT (Lu et al., 2019)	61.8	1.860	4.190
VS (v1) (Dey et al., 2021)	86.8	1.264	3.072
BERT-FT (Kalra et al., 2020)	89.7	1.230	2.982
ALBEF (Li et al., 2021)	57.6	2.220	4.935
ALBEF (ft. on Flickr30k)	64.2	2.242	5.125
ALBEF (ft. on MSCOCO)	64.0	2.002	4.651
LiT (L16L) (Zhai et al., 2022)	64.0	1.849	4.268
CLIP (ViT-B/32) (Radford et al., 2021)	88.1	1.213	2.937
CLIP (ViT-B/16)	92.2	1.123	2.694
CLIP (ViT-L/14@336px)	95.2	1.069	2.547
KAFA (ours)	**97.4**	**1.033**	**2.391**

Table 1: Zero-shot VLM performance on the Pitt Dataset (Hussain et al., 2017) with its official eval protocol (3 positive texts and 12 negative ones for each test image). The best CLIP model already surpasses previous state-of-the-art results (BERT-FT). The size of the data and model used in VLM pretraining have a huge impact on the results. See Sec. 3 for details of the metrics. For completeness, we also include the results of our proposed method (KAFA) here.

4 Challenges in VLM Adaptations to Image Ads and An Intuitive Solution

With CLIP as the clear champion, we further study VLM adaptations for image ad understanding using the best CLIP model (ViT-L/14@336px) as the backbone. We aim to enable better performance for image ad understanding by better adapting pre-trained VLMs to the image ad domain with the help of additional information such as scene-texts extracted from the image.

4.1 The Issue of Overfitting and High Computation Complexity

We find two practical challenges in adapting pre-trained VLMs to the image ads, *first*, the overfitting issue in fine-tuning due to limited image ads and the lack of a strong supervision signal, and *second*, the high computation burden caused by solutions

to the previous challenge.

Annotations of image ads are hard to obtain in general (Akula et al., 2022), making it common to only have limited training data (e.g., the Pitt Dataset only contains 51,223 image-text pairs). This results in VLM's vulnerability to overfitting during adaptation. We find that directly fine-tuning CLIP contrastively on the Pitt Dataset with the symmetric cross-entropy loss (as in the original CLIP paper) gives worse performance than the zero-shot one unless we adopt early stopping and a carefully tuned learning rate schedule. Moreover, as reported in Tab. 1, the best zero-shot performance of CLIP already surpasses the previous state-of-the-art and is very close to 100%, leading to very weak supervision signals for vanilla fine-tuning. We thus need strong training signals. To save GPU memory required by much larger batch sizes, we adopt hard negative mining (Xuan et al., 2020), which selects hard negatives from a very large candidate set as opposed to within the mini-batch.

However, hard negative mining (HNM) strategies usually incur a large computation burden. In fully online hard negative mining (denoted full HNM), for each training image $\mathbf{x}$ and the corresponding texts $\mathbf{y}$, we first rank N_{cand} negative texts $\{y|y \neq \mathbf{y}\}$ sampled from the entire training data according to the online similarity scores (the dot-product score $f_I(\mathbf{x}) \cdot f_T(y)$ computed from the current VLM model), and then we choose the $N_{hard} - 1$ most similar y as the hard negatives. While this essentially constructs a much harder candidate set $\mathcal{C}(\mathbf{x})$, it requires the computation of features of all training texts at every gradient step, which is prohibitively expensive. Existing methods propose to reduce the complexity by keeping a sparsely updated bank of all training text features (memory bank) (Wu et al., 2018) or with the help of a momentum-updated text encoder (MoCo) (He et al., 2020). Nevertheless, we tailor these methods to our setup[1] and find that they perform worse than full HNM. We report the accuracy (%) in Tab. 2 with a harder eval protocol than the official one by using larger numbers (K) of negative samples randomly drawn from the test texts (thus a set of harder negatives).

We believe this is because image ad understanding requires fine-grained information extraction (e.g., the specific brand of a product) and both these two strategies are subject to the destruction of such information as they compute the loss not in a fully online manner. In particular, their text features used for contrastive fine-tuning always come from different VLM encoders, either the past checkpoints or the momentum-updated versions). Although direct fine-tuning with full HNM outperforms the others, it is extremely inefficient and thus impractical.

[1]We use these methods to compute similarity scores but still only select the hardest negatives for fine-tuning to save GPU memory (the purpose of HNM).

Number of Candidates K	20	100	500	1000
Zero-shot	91.7	80.7	64.4	56.5
Direct FT	92.4	82.2	66.7	59.0
Direct FT + memory bank	92.8	82.9	67.5	60.3
Direct FT + MoCo	93.3	83.8	69.8	62.4
Direct FT + full HNM	93.7	84.6	70.0	62.9

Table 2: Accuracy (%) reported with different sizes (K) of the candidate set on the test set of the Pitt Dataset. The larger K means harder negative samples. Zero-shot is the zero-shot performance of the best CLIP model. FT means fine-tuning the best CLIP model.

4.2 Feature Adaptation as the Solution

We propose a simple and intuitive solution, attention-based feature adaptors, that both handle the aforementioned issues during adaptations and enable incorporating additional information (e.g., scene-texts) for better image ad understanding.

Feature adapters are recently proposed (Gao et al., 2021; Zhang et al., 2021; Pantazis et al., 2022; Zhao et al., 2022) as a line of very efficient adaptation strategies of VLMs. They freeze the weights of the pretrained VLMs, pre-compute features using their encoders, and use additional lightweight adapter networks to process these features. As a result, on-the-fly feature computation over a massive candidate set becomes computationally feasible and so is the fully online hard negative mining, since we only compute the adapted features online via a lightweight network. More efficiently, we can set the text adapter to an identity function (i.e., only use adapters for image features).

More importantly, feature adapters are suitable for fusing info from multiple sources. While previous feature adapters are mostly designed for image classification, we consider it as a strategy to aggregate multiple input branches (of potentially different modalities). For instance, previous methods for image ad understanding, such as VS (Dey et al., 2021), utilize scene-texts extracted from images (by OCR) to enhance its performance. Similarly, we can extract text features from scene-texts using a VLM's text encoder and merge them with the

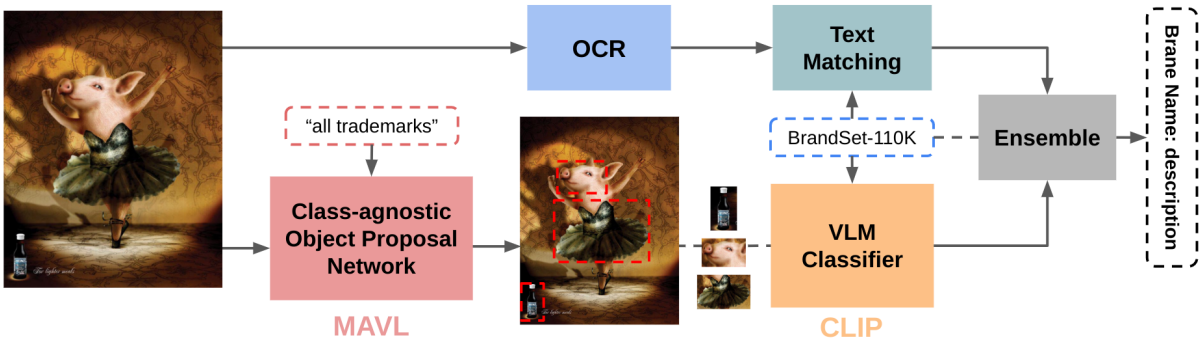

Figure 3: Illustration of our brand understanding module that is an ensemble of text-matching and vision-based recognition. Given an input image ad, we use MAVL to propose regions by prompting "all trademarks" and retrieve entries in BrandSet-110K over the regions with CLIP. We aggregate the predictions across regions and the text-matching results to generate the final output via some simple rules (see details in Appendix B).

image features extracted by the image encoder (of the same VLM) via a feature adapter. In doing so, we obtain a better representation of image ads.

Specifically, we propose to adopt one layer of multi-head attention (Vaswani et al., 2017) as our feature adapter design, similar to the Tiny-Attention Adapter (Zhao et al., 2022). Here the input sequence to the attention layer varies by modalities (brand, scene-texts and image, as in Fig. 4) instead of temporally or spatially as commonly in Transformers. By the nature of alignment-based VLMs, all information (whether in the text format as the scene-texts or the visual elements) are embedded as vectors and lie in a shared semantic space. We then utilize this property and fuse complementary information (e.g., image features and scene-text features) into one feature. Moreover, we append a linear layer after the attention features and equip it with a residual connection. Let us use the notation in previous sections and further denote x_{st} as the scene-texts extracted from the image $\mathbf{x}$ (by Google OCR APIs). Then our adapter is represented as

$$f^{att}(\mathbf{x}) = \mathsf{n}(f_I(\mathbf{x}) + \mathcal{A}[f_I(\mathbf{x}), f_T(x_{st}), ...][0])$$

where $\mathsf{n}(\cdot)$ is a normalization function and $\mathcal{A}$ is multi-head attention (we leave room for other input branches by leaving "..." here). Note that we do not use any adapter for the text descriptions of images (the labels of image ads), which further reduces the computation complexity as now we only need to compute and cache all text features in the training set once and for all during full HNM.

In comparison, we also evaluate the popular CLIP-Adapter (Gao et al., 2021) as a strong baseline, which we tailor to our setup by training three 2-layer residual MLPs. Please see the Appendix for implementation details. As reported in Tab. 3, our proposal of using an attention-based adapter (denoted KAFA w/o K) utilizes VLM features well by aligning multimodal features already in the same semantic space and outperforms CLIP-Adapter. While other existing work (Shen et al., 2021; Gui et al., 2021) merges multiple branches of information by leveraging foundation models, they rely on large encoder-decoder networks that are computationally intensive and might not work well with limited training data as in our case.

5 Improving Image Ad Understanding with External Knowledge

To further improve image ad understanding, we propose to leverage external knowledge of real-world entities, namely product and brand information. The major focus of advertisements is to promote brand awareness (Macdonald et al., 2003). Sometimes brand information is even a necessity to interpret ads correctly since it eliminates ambiguities and gives visual cues to the audiences (e.g., the ad for a cleaning product in Fig. 2). It is then natural to empower feature adapters introduced previously with a brand understanding module that extracts brand information from images. Here we present our training-free brand understanding module that considerably exploits VLMs.

5.1 Brand Understanding Module

Extracting brand information from an image is very challenging due to the sheer scale of brands in the real world. Existing published work (Su et al., 2018; Li et al., 2022a) and even commercial APIs tend to fall short of a good coverage. To solve this

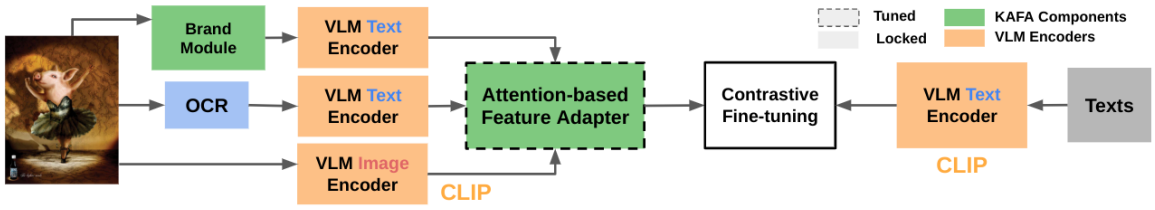

Figure 4: The overall training pipeline of our proposed KAFA, where three branches of information are fed into the attention-based feature adapter, the only neural module free in the fine-tuning process. We leverage VLM encoders for both sides of the contrastive fine-tuning.

issue, we construct a knowledge base that covers brands much better than existing datasets. Our knowledge base has the format, *KFC*: *KFC is a fast food chain*, with around 110k entries covering names of brands, companies, organizations and others appearing in image ads. We call this dataset BrandSet-110K (see details in Appendix B).

Next, we take an ensemble approach to detect and retrieve relevant brand entries from BrandSet-110K given an image ad. On one hand, we retrieve brands by performing string matching over all names in BandSet-110K using the scene-texts extracted by OCR from the image. On the other hand, in case of OCR failures, no detection (some logos have no texts), or multiple detected entries (potentially false positives as most image ads promote only one brand at a time), we use a more powerful vision-based module. Specifically, we adopt MAVL (Maaz et al., 2022), a state-of-the-art VLM, to propose object regions according to the text prompt "all trademarks", We then use the best CLIP model to perform region classification based on a set of carefully engineered prompts. And then, we select the best entries in BrandSet-110K according to the proposed regions. We finally use some simple rules to combine the retrieved results from text-matching and the vision-based module, as in Fig. 3 (see details in the Appendix).

Overall, our brand understanding module is training-free, covers much more entities than previously published work, and even outperforms some commercial logo detection APIs by evaluation on a small validation set, as reported in Tab. 4

5.2 Overall Pipeline and Final Results

Combining with our proposed brand understanding module, we illustrate our overall pipeline in Fig. 4 and call this approach knowledge-augmented feature adaptation (KAFA). In Tab. 3, we demonstrate that KAFA achieves substantial improvements in image ad understanding over the VLM baseline

Method	Inputs	20	100	500	1000
Zero-shot	I	91.7	80.7	64.4	56.5
Direct FT + full HNM	I	93.7	84.6	70.0	62.9
CLIP-Adapter	I+ST	93.9	85.0	70.2	62.8
KAFA w/o K	I+ST	95.0	86.8	72.7	65.1
KAFA w/o ST	I+K	94.7	86.5	72.3	64.5
KAFA (ours)	I+ST+K	**95.6**	**87.7**	**73.9**	**66.0**

Table 3: Accuracy (%) reported on the Pitt Dataset. KAFA (our proposed attention-based adapter with external knowledge) achieves the best results compared to other approaches and the versions with fewer inputs (K = brand knowledge, ST = scene-texts, I = image). Note: "Direct FT + full HN" is extremely inefficient.

	Acc (%)		Acc (%)
VLM-based (ALBEF)	14.5	Text-matching	36.0
VLM-based (LiT)	29.0	Google Cloud API	42.0
VLM-based (CLIP)	64.4	Combined (Text + CLIP)	**66.6**

Table 4: Brand recognition accuracy on ~600 validation image ads. It justifies our brand understanding module and further verifies that models better at recognizing brands are better at image ad understanding.

and consistently outperforms other ablation versions with fewer inputs, justifying that our proposed brand understanding module helps to further improve image ad understanding. We present an example in Fig. 1 to illustrate the improvement of our method over the baseline, where for better display we only show 2 negative text descriptions. See more examples in Appendix G.

6 Additional Analysis

6.1 Hard Negative Samples in Evaluations

We report our main results with a harder eval protocol than the official one. In fact, it is a challenge to perform effective evaluations in retrieval tasks (Akula et al., 2020). While we need **hard** negatives to better reflect the capabilities of a model, usually by increasing the size of the candidate set, we also want those hard negatives to be real **negatives**. As illustrated in Fig. 5 (right), two companies can have two different image ads that share a very similar

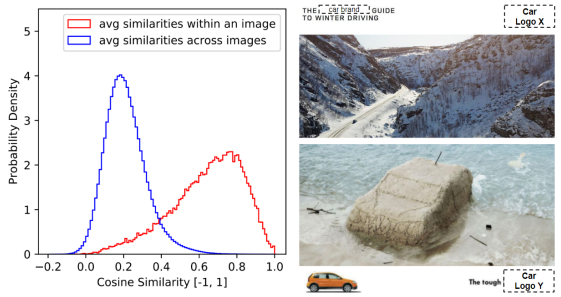

Figure 5: (**Left**) Similarity distributions of texts of the same and across images. Both are spread out with no easy cutoff threshold to sample hard negatives. (**Right**) Two different ads share the same message "I should drive this car because it can drive anywhere", exemplifying the difficulty of sampling hard negative samples.

message. Hence, given an image, simply using a text of another as the negative might not work.

There is no easy solution. We can use a generic sentence encoder to measure similarities among different texts in (Hussain et al., 2017) and only sample texts that are semantically different from the target one (the ground truth) as negatives. We adopt a strong sentence encoder (publicly available here) based on MiniLM (Wang et al., 2020) to measure semantic similarities. We compute similarities among descriptions of the same ad and those across different ads. The similarity distributions are spread out, as demonstrated in Fig. 5 (left), without easy cutoff thresholds to make negative samples both hard and truly negative. Instead, we propose to use several different sizes K of the candidate set with $K = 20, 100, 500, 1000$. For each image in the Pitt Dataset (Hussain et al., 2017), we randomly choose a text from the ground truth and uniformly sample $K - 1$ negatives from other images (harder negatives with larger K).

While most existing methods evaluate (Hussain et al., 2017) with the official evaluation protocol (for ease of comparison we also provide results by this protocol in Tab. 1), it suffers from the lack of hard negatives. Each image ad comes with only 15 randomly sampled candidate texts including 3 positives, giving a random model a 20% accuracy. Moreover, negatives are easy as they tend to be semantically distinct from the positives, making it hard to examine a model at finer levels. We provide examples to compare negatives sampled in our protocol and in the official one in Appendix E.

6.2 Data Leakage Regarding VLMs

The CLIP (Radford et al., 2021) model we use in our experiments was pre-trained on a tremendous

Figure 6: An evaluation image and a found one in LAION-400M. As a reference, this image's caption reads: I should drink "Brand Name" because it'll give me a recharge of energy.

amount (400M) of image-text pairs on the Internet. A concern is that there might be data leakage, i.e., the pre-trained VLMs might have already seen images in the evaluation set, leading to inflated results. We perform an analysis to conclude that this is unlikely the case. We manually inspect images in the LAION-400M dataset (Schuhmann et al., 2021) that are semantically similar to a set of randomly sampled 100 eval image-text pairs. While the dataset used to train CLIP is not publicly released, LAION-400M is a very close one with a similar scale of data filtered by the CLIP model. Specifically, for each of the 100 random samples, we use the open-sourced CLIP-retrieval tool (here) to find the closest images from LAION-400M indexed by both the sample text and image. We do not find any substantially overlapped content or near duplicates (see Fig. 6 as an example). Moreover, our proposed method achieves significant performance improvement over the VLM baseline and both are based on the same CLIP model. Therefore, data leakage is less of a concern.

7 Conclusion

In this paper, we study the adaptation of pretrained alignment-based VLMs for the challenging image ad understanding task. We benchmark and reveal practical challenges in adapting VLMs, propose a simple and intuitive (yet effective) strategy for feature adaptations, and further improve image ad understanding with external brand knowledge. While we mainly focus on the image-to-text retrieval task for its simplicity, we believe further studies can extend it to directly generating text descriptions given image ads or even generating image ads given the descriptions. We hope our study draws more attention to image ad understanding that are relevant to the advertising industry and provide insights for a broader machine learning community.

Limitations

The data from the Pitt Dataset (Hussain et al., 2017), while useful for our paper, contains many images and annotations that may perpetuate harmful stereotypes according to sensitive characteristics such as gender and carry the risk of amplification by machine learning models. We plan to collaborate with AI robustness researchers to identify such examples and develop methods for improving ML models in terms of robustness and reliability.

References

Reneh Abokhoza, Sherehan Hamdalla Mohamed, and Sumit Narula. 2019. How advertising reflects culture and values: A qualitative analysis study. *Journal of Content, Community and Communication*, 10(9):3.

Arjun Akula, Spandana Gella, Yaser Al-Onaizan, Songchun Zhu, and Siva Reddy. 2020. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6555–6565.

Arjun R Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T Freeman, et al. 2022. Metaclue: Towards comprehensive visual metaphors research. *arXiv preprint arXiv:2212.09898*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Lydia B Chilton, Savvas Petridis, and Maneesh Agrawala. 2019. Visiblends: A flexible workflow for visual blends. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Arka Ujjal Dey, Suman K Ghosh, Ernest Valveny, and Gaurav Harit. 2021. Beyond visual semantics: Exploring the role of scene text in image understanding. *Pattern Recognition Letters*, 149:164–171.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*.

François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. 2020. Conceptbert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 489–498.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2021. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1705–1715.

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 *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Kanika Kalra, Bhargav Kurma, Silpa Vadakkeeveetil Sreelatha, Manasi Patwardhan, and Shirish Karande. 2020. Understanding advertisements with bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7542–7547.

Su Young Kim, Hyeonjin Park, Kyuyong Shin, and Kyung-Min Kim. 2022. Ask me what you need: Product retrieval using knowledge from gpt-3. *arXiv preprint arXiv:2207.02516*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Chenge Li, István Fehérvári, Xiaonan Zhao, Ives Macedo, and Srikar Appalaraju. 2022a. Seetek: Very large-scale open-set logo recognition with text-aware metric learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2544–2553.

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. *Advances in neural information processing systems*, 34:9694–9705.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022b. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. 2022. Class-agnostic object detection with multi-modal transformer. In *The European Conference on Computer Vision. Springer*.

Emma Macdonald, Byron Sharp, et al. 2003. *Management perceptions of the importance of brand awareness as an indication of advertising effectiveness*. Ph.D. thesis, Massey University, Department of Marketing.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Omiros Pantazis, Gabriel Brostow, Kate Jones, and Oisin Mac Aodha. 2022. Svl-adapter: Self-supervised adapter for vision-language pretrained models. *arXiv preprint arXiv:2210.03794*.

Savvas Petridis and Lydia B Chilton. 2019. Human errors in interpreting visual metaphor. In *Proceedings of the 2019 on Creativity and Cognition*, pages 187–197.

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. 2021. Combined scaling for open-vocabulary image classification. *arXiv preprint arXiv: 2111.10050*.

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 *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

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

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2021. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. 2022. Clip models are few-shot learners: Empirical studies on vqa and visual entailment. *arXiv preprint arXiv:2203.07190*.

Hang Su, Shaogang Gong, and Xiatian Zhu. 2018. Scalable deep learning logo detection. *arXiv preprint arXiv:1803.11417*.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. 2022. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2015. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742.

Hong Xuan, Abby Stylianou, Xiaotong Liu, and Robert Pless. 2020. Hard negative examples are hard, but useful. In *European Conference on Computer Vision*, pages 126–142. Springer.

Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2021. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*.

Keren Ye and Adriana Kovashka. 2018. Advise: Symbolism and external knowledge for decoding advertisements. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 837–855.

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

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

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133.

Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2021. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*.

Hongyu Zhao, Hao Tan, and Hongyuan Mei. 2022. Tiny-attention adapter: Contexts are more important than the number of parameters. *arXiv preprint arXiv:2211.01979*.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

A Scene-text Extraction by OCR

In our paper, we use scene-texts as one of the inputs for experiments in Pitt Dataset (Hussain et al., 2017). We use the Google Cloud OCR API (link) to extract all text tokens, which are grouped by paragraphs by the API. We then group paragraphs into blocks by simple heuristic rules (e.g., two consecutive paragraphs with similar font sizes should be considered in the same block) and then filter out those blocks with an average prediction confidence score (provided by the API) less than 0.7.

B Brand Recognition

B.1 BrandSet-110K

We construct BrandSet-110K by first compiling entries from public websites. Specifically, for the list of topics (such as automobiles and healthcare) in the Pitt Dataset (Hussain et al., 2017), we Google with the query "Top XX brands/companies in Y" to obtain a list of thousands of common brands, organizations, etc., denote source I. We further scrape the Google Knowledge Graph Search API (link) to find a much larger list of named entities, denoted source II, whose categories fall into "brands", "companies", etc., where each entry comes with a one-paragraph description. Since results from the Knowledge Graph (KG) is a little bit noisy and might miss some popular entities, we rely on source I to make sure that the most prevalent entities appearing in our commercial world are included in our dataset. We then query entries from source I in KG to also obtain the descriptions. If such entries are not found in KG, we simply use the descriptions "X is a brand name in the industry of Y". Together with source II, we obtain a raw combined knowledge base. Then we filter out those entries that are common English words (if the entry appears in the English dictionary (link) or a word set from NLTK (link)). We do so to remove entries such as "Everyday", which will result in too many false positives during brand detection. We also remove entries consisting of a single character. Eventually, we end up with around 110K entries, i.e., name and description pairs.

Since the descriptions returned by KG can be quite long, we further use a learning-based sentence parser to only select the very first sentence of the description (usually in the format of "X is a brand/company/org in the industry of Y with Z features"). We use this API (link) from Hugging Face (Wolf et al., 2019), which is based on spaCy.

B.2 Brand Recognition by Text-Matching

The text-based brand recognition module essentially performs text matching to exhaustively search over all entries in BrandSet-110K given the scene-texts extracted by OCR. For each name in BrandSet-110K that is larger than 6 characters, we match the text in a case-insensitive manner; otherwise, we match it case-sensitively to reduce false positives. A name is set to be matched in a scene-text if it is a phrase of the text ("abc" is matched in "abc def" but not in "abcdef".) When doing ablation studies of evaluating text-matching only performance, in case of multiple predictions we randomly select one as the output.

B.3 Vision-based Brand Recognition

The vision-based brand recognition module handles situations where the text-based one fails (when texts are too small or blurred or artistic for OCR to work; or when logos are purely graphic). The vision-based module is a pipeline of several steps. The class-agnostic region proposal (we use the best model in MAVL (Maaz et al., 2022), a state-of-the-art model) is adopted to generate candidate regions that contain brand logos or visual elements revealing brand information. We choose "all trademarks" as the best prompt with other candidates such as:

- "all small objects", "all brand logos",
- "all brand icons", "all brands", "all logos"

After the region proposal, we use the best CLIP (Radford et al., 2021) model (its visual encoder) to compute the region features. We include the entire image as an extra proposed region. Then we use the text features (via the CLIP text encoder) of the following 6 prompts to find the best entry in BrandSet-110K. Namely

- "A brand logo of X", "A logo of X",
- "A trademark of X","A brand logo of X. Y",
- "A logo of X. Y", "A trademark of X. Y"

where X is the name and Y is the corresponding description in BrandSet-110K. We first average dot products of the region features and brand features across all 6 prompts. We then find two candidates: (1) the name X with the largest predicted scores among all names and all regions of an image and

(2) the name X with the largest predicted scores averaged across all regions among all names that are champions in at least one region. Our final output is chosen by the higher value of the dot products of the global image feature and the two text features of the prompt "'An advertisement of X" (we select this prompt after another minor prompt engineering process).

B.4 Ensemble of Text-matching and Vision-based Brand Recognition

We use simple heuristic rules to ensemble the text-matching results and the vision-based ones. Specifically, if there is no name detected from text-matching, we return the vision-based result; if there is only one name detected from text-matching, we return the text-based result; if more than one name is detected from text-matching, we select the name from detection of both text and vision-based modules by the highest value of the dot product of the global image feature and the text features of "'An advertisement of X". The ensemble module finally returns the single name and the corresponding description in BrandSet-110K.

C Network Architecture of Attention-based Feature Adapter

We adopt a very lightweight network for feature adaptation. For each modality of the inputs (e.g., inputs to KAFA in the Pitt Dataset are three vectors: scene-text features, image features, and brand features), we first add learnable positional embedding (which is used to distinguish between different modalities) and then apply a multi-head attention layer (Vaswani et al., 2017) to obtain a list of vectors; we finally use the first vector (corresponding to the image feature input branch) and add residual connections from the input image feature (before positional embedding) to produce the final output (with normalization). To make things clearer, we also provide the pseudocode.

```
import torch.nn.Parameter as param
import torch.nn.functional as F

# args is a list of input features
# e.g., [img_fs, scene_text_fs, brand_fs]

pos_emb_list = []
for _ in range(n_input):
    pos_emb_list.append(
        param(torch.zeros([input_d])))
attn = torch.nn.MultiheadAttention(
    embed_dim=input_d,
    num_heads=8,
    batch_first=True)
inputs = []
for i in range(n_input):
    inputs.append(
        args[i] + pos_emb_list[i])
x = torch.stack(inputs, 1)
x, _ = attn(x, x, x, need_weights=False)
# The first is the image features.
x = x[:, 0] + args[0]
x = F.normalize(x, dim=-1)
```

D Data Cleaning on Pitt Dataset

We perform data cleaning on both the training and evaluation data of the Pitt Dataset (when evaluated using the official evaluation protocol, whose issue is discussed in the main paper, we stick to the raw evaluation set). For every text in the dataset (the response to the "what should I do according to the ad and why" question), we remove invalid ones (e.g., "I don't know", "not an ad", "not sure"), fix typos (e.g., "becasue", "becaues"), and remove those without answering the "why" question. Furthermore, we filter out texts that do not mention nouns or only have nouns that are not very informative (we compile a list of non-informative nouns appearing frequently in the dataset, such as "product", "thing" and "vendor"). This step is to remove non-specific texts such as "I should buy this product because ...". In the end, we randomly select one text (with a fixed random seed) as the ground truth of its image. If an image has all its texts removed by data cleaning, we remove the image from the dataset. We find such images constituting less than 3% of all images.

E Hard vs. Easy Negatives for Evaluation in Pitt Dataset

Here we explain why we use larger number of candidates K during evaluation. Model evaluation for cross-modal retrieval is challenging (Akula et al., 2020). The official evaluation protocol in Pitt Dataset suffers from the issue that it lacks hard negatives to fully reflect the perception and reasoning capability of the models. Each image in the protocol has 3 positive texts and only 12 negative ones, giving a random guess model a 20% accuracy. On the contrary, increasing the number of candidates in our evaluation protocol as introduced in the main paper effectively yields harder negatives. For instance, for the image ad in Fig. 7 whose ground truth is "I should buy a Brand A camera because it will help me create", if we set the number of candidates to be 10 (i.e., 9 negatives), the best

CLIP model makes the correct selection with all easy negatives, among which the most confusing ones are

- "I should drink Brand B because it de-ages you"
- "I should not drown in my decision because bad choices will keep you under"
- "I should buy this bag because it is resealable"

If we set 50 total candidates (i.e., 49 negatives), again the CLIP baseline predicts correctly with the most confusing ones still being relatively easy negatives:

- "I should use Brand C cosmetics because it makes you beautiful"
- "I should buy Brand D products because they are reliable"
- "I should see a movie because it's fun"

For a larger number (e.g., 100 total candidates), the CLIP model starts to make mistakes, with hard negatives such as

- "I should use Brand E makeup because it will make me more seductive"
- "I should buy Brand F makeup because it will make me beautiful"
- "I should buy this makeup because it will make me shine"

Notice that for privacy reasons, all brand names in this example are anonymized.

F Training Details

F.1 Direct Fine-tuning of CLIP

We fine-tune the best CLIP model on the training images of Pitt Dataset with a batch size of 8, symmetric cross-entropy loss (the one used in the original paper of CLIP) and the Adam optimizer (Kingma and Ba, 2014) with weight decay of $1e-4$. We set other parameters of Adam as in the original implementation of CLIP. We find that using a very small learning rate (e.g., $1e-7$) is necessary for fine-tuning CLIP on Pitt Dataset; otherwise, the CLIP model can overfit easily. For the same reason, we adopt early stopping and only fine-tune the model for a maximum of 4 epochs. We leave the details in the next section for the fine-tuning version with online hard negative mining (very computationally intensive as suggested in the main paper).

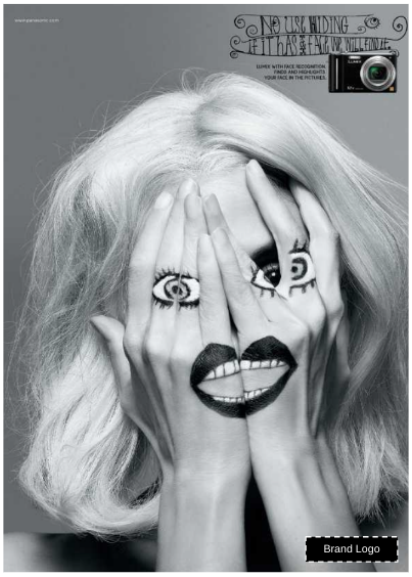

Figure 7: An example to illustrate the issue of easy negative samples in evaluation.

F.2 Fully Online Hard Negative Mining (full HNM)

When performing hard negative mining during training, for each image in a mini-batch, we first compute the VLM features of a large number of randomly sampled negative texts (in our experiments we find 1000 to be large enough; while a larger number can marginally improve the final performance but it incurs a larger computation burden), then we compute the dot products of the current image feature and all these sampled text features, and finally, we rank the dot products and select the top $N-1$ negatives to be included in computing the gradients of the loss (we find $N=8$ to be effective). We use the asymmetric version of the cross-entropy loss (i.e., the normal one) compared to the asymmetric version in CLIP pre-training since the number of negatives per image does not equal the batch size when HNM is adopted. We reduce the batch size to 4 whenever with online HNM so that directly fine-tuning the largest CLIP model is viable with a single V100 Nvidia GPU. We still apply the learnable "logit scale" parameter in CLIP pre-training which effectively makes contrastive learning more stable.

For full HNM, if we directly fine-tune the CLIP model, we need to compute text features of all texts in the training set in every gradient step. While this is computationally prohibitive, we adopt the feature adapter strategy and thus cache all the text features once and do not update the text encoder

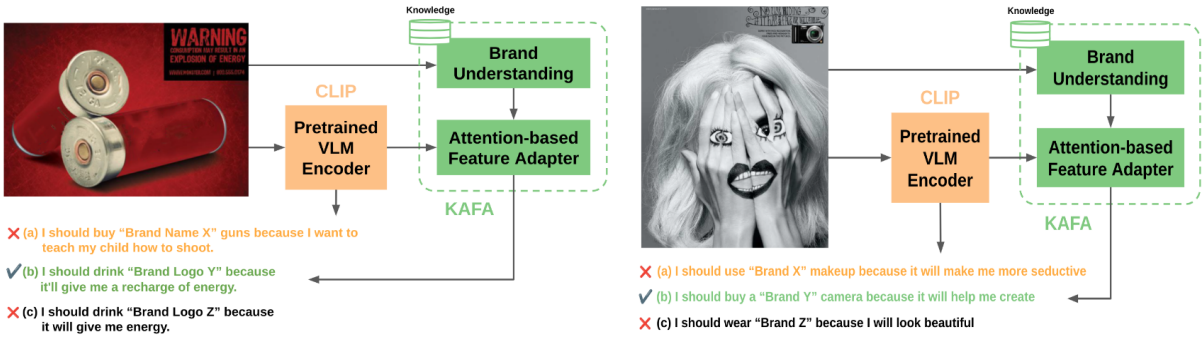

Figure 8: Additional examples that demonstrate KAFA's improvements over the VLM baseline.

and the text features during fine-tuning.

F.3 More Ablation Studies

In our experiments presented in the main paper (specifically in Tab. 1), we have justified the use of online HNM, the additional inputs (scene-text and brand information) to the feature adaptation, and the advantages of the attention-based adapter over the baseline adapter. We also perform experiments on several variants of the attention-based feature adapter and find that either using more than one attention layer or adding layer norm & additional linear projection as in the encoder-decoder Transformer (Vaswani et al., 2017) make the model more vulnerable to overfitting.

F.4 Additional Details of Feature Adapters

For feature adapters (CLIP-Adapter and KAFA), we use the full HNM for fine-tuning as discussed in the previous section. We use the same training setup as that of "Direct ft + HMN" except for the additional input branches. For CLIP-Adapter, we tailor it to our setup by training three 2-layer residual MLPs. Specifically, let as denote them as g_I^{mlp}, g_T^{mlp} and h^{mlp}, built on top of the image and text features extracted by VLMs, and a mixture of these features, respectively. The adapted feature for $\mathbf{x}$ becomes

$$f_I^{mlp}(x) = \mathsf{n}(f_I(x) + g_I^{mlp}(f_I(x)))$$
$$f_T^{mlp}(x_{st}) = \mathsf{n}(f_T(x_{st}) + g_T^{mlp}(f_T(x_{st})))$$
$$f^{mlp}(x) = \mathsf{n}(h^{mlp}(\mathsf{cat}[f_I^{mlp}(x), f_T^{mlp}(x_{st}), ...]))$$

where cat is concatenation. Here we omit the adapted feature for text label y. And the adapted feature for the text label y becomes

$$f_T^{mlp}(y) = \mathsf{n}(f_T(y) + g_T^{mlp}(f_T(y)))$$

which is used during full HNM for fine-tuning.

For fine-tuning of both CLIP-Adapter and KAFA, we find a much larger learning rate (i.e., $1e-4$) to be effective and train the model similarly with early stopping and a maximum of 10 epochs. We find it helpful to stabilize training by adding an additional regularization loss to keep the feature adapter's output close to the VLM image features. Specifically, we add the negative of dot products between the two (averaged over all data points in the mini-batch) to the overall training objective. For this regularization term, we use a coefficient of 5 in all our experiments in the Pitt Dataset.

G Additional Examples

We present 2 additional examples in Fig. 8 to illustrate the improvement of our method over the baseline. Again, we only show 2 negative text descriptions for better display, and we anonymize all brand info.