

Don't Buy it! Reassessing the Ad Understanding Abilities of Contrastive Multimodal Models

Anna Bavaresco, Alberto Testoni, Raquel Fernández

Institute for Logic, Language and Computation

University of Amsterdam

{a.bavaresco, a.testoni, raquel.fernandez}@uva.nl

Abstract

Image-based advertisements are complex multimodal stimuli that often contain unusual visual elements and figurative language. Previous research on automatic ad understanding has reported impressive zero-shot accuracy of contrastive vision-and-language models (VLMs) on an ad-explanation retrieval task. Here, we examine the original task setup and show that contrastive VLMs can solve it by exploiting grounding heuristics. To control for this confound, we introduce TRADE, a new evaluation test set with adversarial grounded explanations. While these explanations look implausible to humans, we show that they “fool” four different contrastive VLMs. Our findings highlight the need for an improved operationalisation of automatic ad understanding that truly evaluates VLMs’ multimodal reasoning abilities. We make our code and TRADE available at <https://github.com/dmg-illc/trade>.

1 Introduction

Image-based advertisement is not only a crucial component of marketing campaigns, but also an interesting example of sophisticated multimodal communication. Ads often feature unusual visual elements (e.g., objects that are non-photorealistic, outside of their usual context, atypical, etc.) or examples of figurative language (e.g., metaphors, allegories, play on words, etc.) designed to make a long-lasting impression on the viewer. Figure 1 provides an example of an ad with a non-photorealistic object (a whale made of wires) that, as the text suggests, is used to convey a complex metaphorical message about the product (i.e., a wireless device).

These elaborate uses of images and text make automatic ad understanding a challenging task requiring multiple non-trivial abilities, e.g., object detection, scene-text extraction, figurative language understanding, and complex image-text integration. Ad understanding was first proposed as a deep-

learning task by Hussain et al. (2017), who introduced the Pitt Ads dataset, consisting of image-based ads along with explanations capturing their underlying message (e.g., *I should purchase this stereo system because wireless is less messy*). This dataset was then used in a retrieval-based challenge requiring to identify a plausible explanation for an ad within a set of possible candidates.¹

Early work on this task has employed ensemble predictors (Hussain et al., 2017; Ye and Kovashka, 2018) and graph neural networks (Dey et al., 2021) that were designed and trained *ad hoc*. More recently, the development of large vision-and-language models (VLMs) pretrained with image-text matching (ITM) objectives has opened the possibility of performing the task in zero-shot, i.e. by using an off-the-shelf model instead of training one from scratch. Following this approach, Jia et al. (2023) tested multiple VLMs (ALBEF, Li et al. 2021; CLIP, Radford et al. 2021; and LiT, Zhai et al. 2022) on the task by computing image-text alignment scores between ads and their possible explanations. They observed an excellent zero-shot performance for all models, documenting an accuracy of 95.2% for CLIP.

While the results reported by Jia et al. (2023) seem to suggest that the tested models developed the reasoning abilities necessary to succeed at ad understanding, we note that this conclusion is in contrast with a great deal of existing work. Extensive research investigating whether VLMs develop reasoning skills as a result of their contrastive ITM pretraining has exposed several weaknesses of these models. They have been shown to be limited in their abilities to identify noun mismatches in image captions (Shekhar et al., 2017), reason compositionally (Thrush et al., 2022), capture spatial relations (Liu et al., 2023), understand verbs (in-

¹<https://eval.ai/web/challenges/challenge-page/86/overview>

Text on the ad: Wires are under extinction. DVD theater [brand name]. Now wireless.



Original task setup

1. I should get a [wbn] bike because they have been around for a while
2. *I should buy a [brand name] because I will not need the wires*
3. *I should use wireless instead of wires because it will help reduce waste in the world*
4. I should not eat meat because it supports the killing of animals
5. I should not wear fur because it kills animals
6. I should buy ice cream because it's on sale for the company being in business for 16 years
7. I should fund [wbn] because we should take back control
8. *I should purchase this stereo system because wireless is less messy*

Our task setup in TRADE

1. I should use wires because they are like whales risking extinction
2. *I should use wireless instead of wires because it will help reduce waste in the world*
3. I should use caution when throwing away cables because whales risk extinction

Figure 1: An example of the ad explanation retrieval task with the original setup vs. our new setup. The matching explanations are marked in italics. In the original setup, negatives are randomly sampled (5 out of 12 are shown for conciseness); in our setup, negatives are carefully curated to be textually and visually grounded in the ad but, at the same time, clearly incompatible with it. Brand names and logos are edited out in the examples present in this paper for presentation purposes, but are in fact visible in both task setups ([wbn] stands for “wrong brand name”).

stead of just nouns) (Hendricks and Nematzadeh, 2021), and handle various linguistic phenomena (Parcalabescu et al., 2022) and basic constructions (Chen et al., 2023).

Importantly, this line of work focused on a set of traditional visuo-linguistic tasks but not specifically on ad understanding. Here, we ask whether the performance previously documented on the Pitt Ads dataset reflects genuine understanding abilities or is driven by simpler heuristics. We conduct a thorough analysis of the evaluation setup originally proposed to test ad understanding and reveal that it has key flaws, which allow models to exploit grounding heuristics. We introduce a new test set, TRADE (*TRuly ADversarial ad understanding Evaluation*), which controls for the identified issues. Our experiments show that several contrastive models tested zero-shot, including CLIP, perform at chance level on TRADE, while humans excel at the task. More generally, our findings highlight the need to better operationalise ad understanding in order to obtain reliable assessments of VLMs’ multimodal reasoning abilities.

2 A Closer Look at the Evaluation Setup

The Pitt Ads dataset² by Hussain et al. (2017) consists of 64832 ads, each annotated with 3 explanations in English written by 3 different expert annotators. These explanations (in the form *I should <action> because <reason>*) aim at capturing the persuasive message behind the ads. While explanations may be subjective, the intuition behind the image-to-text retrieval task proposed along with

the dataset is that a model which can understand ads should be able to match them with a plausible explanation. Specifically, each ad is paired with 15 messages, 3 positives corresponding to the annotations for that ad and 12 negatives randomly sampled from annotations for different ads. Figure 1 provides an overview of the task setup.

Previous work has hinted at possible limitations of the evaluation setup. Kalra et al. (2020) observed a significant overlap between the text present in the ad and the matching explanations and noticed this was “a major discriminating factor” that their fine-tuned BERT model could exploit. Similarly, Jia et al. (2023) pointed out that the candidate set lacks “hard negatives” and proposed to increase the set size, but could not provide a solution ensuring the negatives were actually hard.

We conduct a quantitative analysis on the original evaluation setup to uncover potential shortcuts that VLMs may be exploiting to solve the task. We hypothesise that the models may take advantage of two factors that do not necessarily reflect ad understanding: simple relationships between (1) the candidate explanations and the text present in the ad (i.e., the degree of *textual grounding* of the explanations) and (2) the entities mentioned in the explanations and those depicted in the image (their degree of *visual grounding*). To test our hypotheses, we define several visual- and textual-grounding scores and check whether they correlate with the CLIP-based alignment score used by Jia et al. (2023) to retrieve the ad explanations.³

²<https://people.cs.pitt.edu/~kovashka/ads/>

³More details on the scores can be found in Appendix A.

Textual-grounding scores are computed between candidate explanations and the text extracted from the ad with Optical Character Recognition (OCR). We calculate (1) *text overlap* as the proportion of content-word lemmas from the explanation that are also present in the OCR-extracted text, and (2) *text similarity* as the cosine similarity between a sentence-level embedding of the explanation and that of the OCR-extracted text, derived with MPNet (Song et al., 2020).

Visual-grounding scores include (1) *object mention* as the proportion of nouns in the candidate explanation that are present in a set of objects we automatically extracted from the image by a ResNet50 model (He et al., 2016), and (2) *caption similarity* as the cosine similarity between the sentence-level embedding of the candidate explanation and the embedding of the ad caption we obtained with BLIP-2 (Li et al., 2023). Our motivation for examining both detected objects and generated captions is driven by the observation that they capture complementary information. More specifically, while detected objects are not mediated by language models, they may often incorporate non-salient objects that people would unlikely mention when describing a picture or contain lexical choices that differ from the human ones. On the other hand, generated captions refer to objects in a more human-like way but, at the same time, may contain hallucinations due to linguistic priors.

We compute the grounding scores and CLIP’s alignment score for the test split of the Pitt Ads dataset, consisting of 12805 samples. As hypothesised, we observe a positive correlation between all our grounding scores and CLIP’s alignment score. All the Spearman’s correlation coefficients are significant ($p \ll 0.001$) and range from 0.14 and 0.61 (see Appendix A for details). In addition, as shown in Table 1 (left), we find that in the original setup the matching explanations are significantly more grounded than the non-matching explanations for each ad. While the elements (OCR text, objects, captions) detected by other models are not necessarily the same as those identified by CLIP, these results suggest that reasonably similar information is indirectly extracted by CLIP and exploited to solve the ad-understanding task. This finding also agrees with results from previous work showing that CLIP develops OCR capabilities and can successfully classify objects (Radford et al., 2021).

Overall, these results indicate that the original

	original setup		TRADE	
	Pos	Neg	Pos	Neg
<i>text overlap</i>	0.21	0.03 *	0.27	0.31 *
<i>text similarity</i>	0.40	0.12 *	0.44	0.42
<i>object mention</i>	0.03	0.01 *	0.02	0.04
<i>caption similarity</i>	0.32	0.11 *	0.34	0.35

Table 1: Average textual- and visual-grounding scores of the matching (Pos) and non-matching (Neg) explanations in the original evaluation setup and in TRADE; statistically significant differences between Pos and Neg marked with * ($p \ll 0.001$, two-sample t-test).

evaluation setup is flawed and that the outstanding zero-shot performance obtained by VLMs on the retrieval task may be due to simple image-text alignment.

3 TRADE: A New Adversarial Test Set

To test the extent to which VLMs capture elaborate visuo-linguistic relationships present in image-based ads beyond image-text alignment, we develop TRADE (*TRuly ADversarial ad understanding Evaluation*), a new diagnostic test set with adversarial negative explanations. TRADE consists of 300 randomly selected ads from the Pitt Ads dataset, each associated with 3 options (1 positive and 2 negatives). Concretely, for each of these ads, we randomly select one valid explanation from the available annotations and create two adversarial negative explanations—see Figures 1 and 2 for examples (more examples in Appendix C). The adversarial explanations were created by 4 expert annotators who were instructed to do their best to come up with non-plausible explanations that nevertheless mention objects and fragments of text present in the image. Annotators were also asked to approximately match the length of the positive explanation when writing these adversarial sentences. Appendix B contains more details about the creation of the adversarial negatives, including the guidelines provided to the annotators.

We validate TRADE in two ways. First, we compute the textual- and visual-grounding scores introduced in Section 2. This shows that in TRADE the gap between positive and negative explanations is radically reduced compared to the original setup, as can be seen in Table 1 (right). Second, we confirm that humans are not affected by the high level of grounding of both positive and negative examples and are able to identify the plausible explanation in

the TRADE samples with an accuracy of 94%.⁴

To allow for a direct comparison with an evaluation setup with random negatives, akin to the original task setup, we also create TRADE-control: a version of TRADE where the two negative explanations per ad are randomly sampled from the explanations for other ads. TRADE-control includes 10 versions created with different random samplings.

TRADE and TRADE-control are publicly available at <https://github.com/dmg-illc/trade> under a Creative Commons Attribution 4.0 International (CC-BY) license.

4 Experiments

We use TRADE to test four contrastive pretrained VLMs zero-shot. Three of these models (CLIP, Radford et al. 2021; ALBEF, Li et al. 2021; and LiT, Zhai et al. 2022) have been shown to achieve high zero-shot performance on the original task setup (Jia et al., 2023). Here we challenge them with TRADE and consider an additional model (ALIGN, Jia et al. 2021).

4.1 Models and Setup

Except for ALBEF, all the models we test encode visual and textual inputs separately and are pretrained with an image-text matching objective. ALBEF has an additional multimodal module, but here we only use its unimodal encoders, which are also pretrained contrastively. A more detailed overview of these VLMs is reported in Appendix E.

All four models allow for the computation of an image-text alignment score, here defined as the dot product between the normalized image embedding and the text embedding of each candidate explanation. As in previous work (Jia et al., 2023), we evaluate the models by computing alignment scores for every ad-explanation pair and consider the explanation yielding the highest alignment score as the model’s retrieved option. We report average accuracy, as (mean) rank is not very informative with only 3 candidates.

4.2 Results

Table 2 shows the performance of the models on TRADE and TRADE-control. All models achieve an accuracy higher than 80% in the control condition, with CLIP reaching 98%. However, the performance of all models in the adversarial setting—

⁴Each of the 300 samples was annotated by two annotators external to the project; more details available in Appendix D.

Model	TRADE	control
CLIP (ViT-L/14@336px)	0.34	0.98 (0.01)
ALIGN (base)	0.28	0.97 (0.01)
LiT (L16L)	0.31	0.82 (0.02)
ALBEF (ft. on Flickr30k)	0.33	0.88 (0.01)

Table 2: Average accuracy on TRADE vs. TRADE-control. The TRADE-control values are averages over 10 random samples, with standard deviation in brackets.

where humans achieve 94% accuracy, cf. Section 3—nears chance level, i.e., 33%. Figure 2 provides an example of model- and human-chosen ad explanations on a TRADE instance. These results provide compelling evidence that the evaluated VLMs rely on visual and textual grounding when retrieving ad explanations. As a result, they can achieve excellent accuracy in an evaluation setting where negatives are poorly grounded, but are easily “fooled” by grounded adversarial distractors that are extremely easy for humans to discard.

To get more insight into the models’ performance, we examine their predictions and observe that, while all models perform equally poorly on TRADE, there are 23 samples (8% of the dataset) for which the four models succeed at identifying the target explanation. An analysis of the explanations correctly retrieved by all models reveals that most of them exhibit grounding scores that are higher than the average scores for matching explanations. Figure 3 visualises this finding.

5 Conclusions

Our work exposes key limitations of the evaluation setup that was previously used to benchmark VLM’s ad understanding abilities. We introduce a new adversarial test set (TRADE) that controls for the identified issues and show that, while humans excel, contrastive VLMs perform at chance level on TRADE. This result has the following implications.

First, it shows that, when processing image-based ads, contrastive VLMs are strongly biased towards textually and visually grounded explanations, regardless of their plausibility. This is in agreement with previous work (Hendricks and Nematzadeh, 2021; Parcalabescu et al., 2022; Thrush et al., 2022; Liu et al., 2023; Chen et al., 2023) and points to the need to use caution when interpreting models’ zero-shot accuracy on “naturalistic” (i.e., non-adversarial) setups as proof that they develop sophisticated reasoning abilities via pretraining.


Ad	TRADE explanations	Chosen by
	1. <i>I should go to [brand name] not only does their food taste great but it also looks good.</i>	Human
	2. I should go to [brand name] because my eyelashes need a new look.	CLIP, ALIGN
	3. I should go to [brand name] because tasty burgers must look like these eyelashes.	ALBEF, LiT

Figure 2: Ad explanations selected by human annotators vs. our tested models for one instance from TRADE. Italic indicates the *matching explanation*. Brands and logos are edited out in the paper examples for presentation purposes but are visible to models and human annotators.

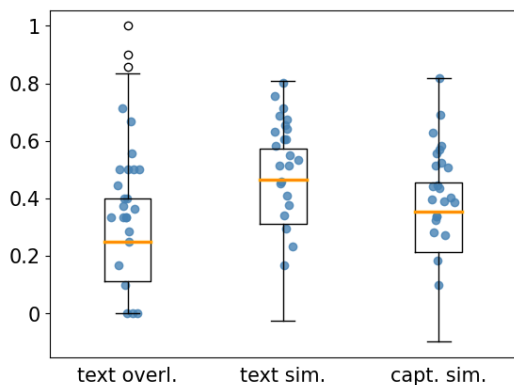


Figure 3: Boxplots summarizing the distribution of grounding scores computed for positive explanations in TRADE. The blue dots indicate the scores for the positive explanations correctly selected by all VLMs. The *object mention* score is not included because its median coincides with the quartiles.

Second, our work highlights issues with the current retrieval-based operationalization of ad understanding as a task to evaluate VLM’s multimodal reasoning abilities. We emphasise that TRADE’s aim is to control for a confound—the grounding gap between positives and negatives—that we identified as crucial when testing a specific type of VLMs, i.e., those pretrained with an ITM objective. However, defining which abilities are necessary to conclude that a model developed a good “understanding” of image-based ads and designing a task that truly evaluates them remain open issues for future research.

Limitations

The current study and previous work have operationalised ad understanding as an ad-explanation retrieval task. In particular, we have focused on testing contrastive pretrained VLMs zero-shot on this task. Consequently, the question of whether

VLMs trained or finetuned on the Pitt Ads dataset would be more robust against our adversarial explanations remains open and could be investigated in the future. Nevertheless, we emphasize that the retrieval-based setup has limitations (e.g., the impossibility of providing task-specific instructions to the models) and may not be the most appropriate to evaluate VLM’s ad understanding skills and their multimodal reasoning abilities more generally. An interesting direction for future research could be to formulate the task differently, e.g., as a generative task. This would solve some issues of the retrieval-based setup, but also posit novel challenges, such as identifying the most effective prompt and defining meaningful protocols to evaluate the generated explanations.

On a methodological note, we highlight that visual and textual alignment are complex constructs that encompass different aspects and can be analysed at different levels of granularity. Therefore, we do not intend our grounding scores as precise and comprehensive metrics, but simply as indicators that can reflect general trends.

Ethical Considerations

TRADE does not introduce new ad-images, but simply links to the existing Pitt Ads dataset along with the set of adversarial explanations we have created. However, it is worth emphasizing that the ads present in Pitt Ads were originally collected by querying Google Images. This posits two ethical concerns.

First, offensive/harmful content or stereotypes may be present in the images, as already pointed out by [Jia et al. \(2023\)](#). To minimise this potential problem when developing TRADE, we made sure the annotators who created our adversarial explanations had the possibility of flagging ads that they deemed inappropriate (they did so a couple

of times). However, we cannot fully guarantee that the ad images used in TRADE are completely free from harmful content. As for the adversarial distractors created for TRADE, we have not systematically examined all of them manually to make sure they do not contain harmful content, but we believe this is very unlikely given the guidelines and the fact that they were created in a very controlled setting partially by us and partially by close colleagues.

The second concern is about the license of the images. The Pitt Ads dataset was released without a license and the curators do not clarify whether the images are copyrighted or not.

Finally, we note that our study does not take into account the personal and cultural factors which may play a substantial role in people’s perception of ads or in the values they associate with certain products. Although TRADE includes only one matching explanation for each ad, we emphasize that we do not intend this as a “ground truth”. We hope that future research on automatic ad understanding will adopt evaluation protocols that reflect a diverse set of possible interpretations.

Acknowledgements

We warmly thank the current members and alumni of the Dialogue Modelling Group (DMG) from the University of Amsterdam for the support they provided at different stages of this project. Special thanks are due to Sandro Pezzelle, who suggested the name ‘TRADE’ for our introduced dataset. Our heartfelt gratitude also goes to the colleagues who assisted us with the creation of TRADE negatives and to the colleagues, friends and partners who kindly volunteered to participate in our experiment to assess human performance on TRADE. The present work was funded by the European Research Council under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 819455).

References

Xinyi Chen, Raquel Fernández, and Sandro Pezzelle. 2023. [The BLA benchmark: Investigating basic language abilities of pre-trained multimodal models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5817–5830, Singapore. Association for Computational Linguistics.

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.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. [Deep residual learning for image recognition](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. [Probing image-language transformers for verb understanding](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online. Association for Computational Linguistics.

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 (CVPR)*, 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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 4904–4916.

Zhiwei Jia, Pradyumna Narayana, Arjun Akula, Garima Pruthi, Hao Su, Sugato Basu, and Varun Jampani. 2023. [KAFA: Rethinking image ad understanding with knowledge-augmented feature adaptation of vision-language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 772–785, Toronto, Canada. Association for Computational Linguistics.

Kanika Kalra, Bhargav Kurma, Silpa Vadakkeveetil 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, Online. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. [BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

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](#).

- In *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705. Curran Associates, Inc.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the 13th European Conference on Computer Vision (ECCV)*, pages 740–755. Springer.
- Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. [Visual spatial reasoning](#). *Transactions of the Association for Computational Linguistics*, 11:635–651.
- Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. [VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics.
- 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. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763.
- Andrey Savchenko, Anton Alekseev, Sejeong Kwon, Elena Tutubalina, Evgeny Myasnikov, and Sergey Nikolenko. 2020. [Ad lingua: Text classification improves symbolism prediction in image advertisements](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1886–1892, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. [FOIL it! find one mismatch between image and language caption](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada. Association for Computational Linguistics.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MpNet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, Red Hook, NY, USA. Curran Associates Inc.
- Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5238–5248.
- 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.
- 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 (CVPR)*, pages 18123–18133.

Appendix

A Grounding Scores

Textual Grounding The textual grounding scores were computed between candidate explanations and the ad OCR-extracted text output by Google Vision API⁵ and made publicly available⁶ by the authors of [Savchenko et al. \(2020\)](#). OCR text was present for 12304 ad-images of the test set and 294 images from TRADE. The *text overlap* score was computed as the proportion of words from the candidate explanation that were also present in the OCR-extracted text. Before computing the overlap, we lemmatized the text and removed stop-words. These preprocessing steps were performed with the NLTK⁷ package.

The *text similarity* score was defined as the cosine similarity between the embedding of the explanation and that of the OCR-extracted text. The embeddings were obtained using the Sentence Transformers⁸ framework. Specifically, we used an MP-Net ([Song et al., 2020](#)) pretrained model, which was indicated as the best-performing one.

Visual Grounding To compute the visual grounding scores, we considered two sources of visual information: the objects identified by an object detector, and the ad-image captions. Our object detector was a ResNet50 model ([He et al., 2016](#)) pretrained on MS COCO ([Lin et al., 2014](#)). We used the implementation from the Detectron2⁹ framework by Facebook. The model detected an average of 3.74 ± 3.85 objects from the Pitt Ads dataset test split and 3.51 ± 3.38 from TRADE. At least one object was detected on 11351 images from the Pitt Ads dataset test split and on all the ads

⁵<https://cloud.google.com/vision/docs/ocr>

⁶https://figshare.com/articles/dataset/OCR_results/6682709

⁷<https://www.nltk.org/>

⁸<https://github.com/UKPLab/sentence-transformers?tab=readme-ov-file>

⁹<https://github.com/facebookresearch/detectron2>

from TRADE (300). Ad captions were obtained using BLIP2 (Li et al., 2023) with OPT 2.7B as language decoder. BLIP2 was used in its Hugging Face implementation.¹⁰

The *object mention* score was computed as the lemmatized nouns in the AR statement that were part of the set of detected objects. The *caption similarity* score, on the other hand, was defined similarly to the *text similarity* score, with the caption in place of the OCR-extracted text.

Correlation with CLIP’s alignment scores We computed Spearman correlations between all the grounding scores and the CLIP-alignment scores for both the Pitt Ads test set and TRADE.

All results are summarized in Tables 3 and 4.

B Creating TRADE

The adversarial negatives were designed by two of the authors and two internal collaborators who volunteered for the task and are all proficient in English. Due to the complexity of this annotation task, we deemed it not suitable for crowdsourcing. The instructions given to the annotators were the following:

1. The sentence should be inconsistent with the image, meaning that it should not be a valid answer to the question “What should you do, according to this ad?”. Keep in mind that the answer should be patently wrong, i.e. it should require very little thinking to figure out it does not match the message of the ad.
2. The sentence should be in the form “I should [action] because [reason]”
3. The verb you use after “should” should be the same as the one from the right sentence. For example, if the right sentence starts with “I should buy”, your wrong annotation cannot start with “I should fly”
4. The sentence should be as grounded as possible, meaning that you should avoid mentioning objects/words that are not present in the ad as much as you can. Please keep this in mind, it is very important!
5. If possible, privilege salient visual elements over non-salient ones. More concretely, try to mention large writings instead of small ones, and big foreground objects instead of small background ones.
6. When describing visual objects, try to be efficient instead of verbose. For example, if an ad depicts a famous man (say, Mr. X) driving a car of a specific brand (say, Brand Y), you should write something like “I should buy Mr. X because he drives a cool Brand Y car” instead of “I should buy a man with short hair and sunglasses because he drives a red four-wheeled vehicle”
7. Please avoid extra-long sentences. Your wrong answers should be approximately the same length as the correct ones. You don’t need to be as strict as to count the exact

¹⁰https://huggingface.co/docs/transformers/model_doc/blip-2

number of words but try to avoid large mismatches (e.g. correct answer being not even one-line long and wrong answer being two lines)

8. Only include the name of brands/celebrities if they are also mentioned in the provided annotation
9. The sentence (e.g., “I should buy this perfume because roses are red and violets are blue”) but it should not be ungrammatical (do not write something like “I should hello world because rainbow”)

Rule 8 was introduced as there is evidence (Goh et al., 2021) that CLIP is sensitive to proper nouns. Therefore, we wanted to avoid our negatives being preferred by the model simply because they contained more detailed information.

Our annotation interface allowed annotators to flag ads in case of:

1. Presence of inappropriate/offensive/harmful content.
2. Low readability of the text.
3. Low image resolution.
4. Being unable to understand the ad (e.g., because the text was not in English).
5. Being unable to create a distractor meeting all the requirements.

C Dataset Examples

Some additional examples of the adversarial explanations we collected are shown in Figure 4 along with their TRADE-control counterparts.

D Human Accuracy on TRADE

To quantify the human accuracy on TRADE, we used the crowdsourcing platform Appen to present participants with the ad along with the question “What should you do according to this ad, and why?” and 3 options, i.e., a matching explanation and two adversarial grounded negatives. After some unsatisfactory pilot experiments where crowdworkers were not able to pass very simple test questions, we established that the task was not suitable for crowdsourcing. Therefore, we recruited 17 participants who volunteered for the task of judging the 300 samples in TRADE. They were not involved in the creation of the adversarial explanations and were informed that their anonymised data would be included in a study about automatic ad understanding. We ensured all annotators were proficient in English. Each question was answered by 2 different participants. They annotated an average of 35 ads each ($std = 14$, $max = 50$, $min = 10$). The mean accuracy calculated over the 600 collected

	Pos	Neg	CLIP-pos	CLIP-neg	Corr
<i>text overlap</i>	0.21	0.03	23.78	12.66	0.28
<i>text similarity</i>	0.4	0.12	23.78	12.66	0.61
<i>object mention</i>	0.03	0.01	23.72	12.74	0.14
<i>caption similarity</i>	0.32	0.11	23.72	12.68	0.53

Table 3: Grounding scores and CLIP-alignment scores for matching (positives) and non-matching (negatives) explanations from the original test set. Two-sample t-tests indicate that all differences between positives and negatives are statistically significant ($p \ll 0.001$). The right-most column reports the Spearman correlations between aggregated (including both positives and negatives) grounding scores and the corresponding CLIP-alignment scores. All the correlation values are statistically significant ($p \ll 0.001$).

	Pos	Neg	CLIP-pos	CLIP-neg	Corr
<i>text overlap</i>	0.27	0.31	24.87	24.42	0.22 ($p = 0$)
<i>text similarity</i>	0.44	0.42	24.87	24.42	0.41 ($p = 0$)
<i>object mention</i>	0.02	0.04	24.84	24.39	0.04 ($p = 0.22$)
<i>caption similarity</i>	0.34	0.35	24.84	24.39	0.3 ($p = 0$)

Table 4: Grounding scores and CLIP-alignment scores for matching (positives) and non-matching (negatives) explanations from TRADE. With the exception of *text overlap*, the differences between grounding scores are not statistically significant ($p \ll 0.001$). All the differences between positive and negative CLIP-alignment scores are also non-significant.


	<p>Text: Quick. Slow. Want help? Phone the smokeline on 0800 84 84 84. You can do it. We can help.</p> <p>Explanation: <i>I should stop smoking because it is slowly killing me</i></p> <p>TRADE distractors:</p> <ul style="list-style-type: none"> • I should stop smoking because I want a quick help • I should stop smoking because bullets are slow <p>TRADE-control distractors:</p> <ul style="list-style-type: none"> • I should wear [clothing brand] because it is natural. • I should buy [makeup brand] makeup because it has bold lipstick colours
	<p>Text: [brand name] Up your game. Lane Carico [brand name] elite athlete.</p> <p>Explanation: <i>I should buy these shoes because they will help you perform sports really well</i></p> <p>TRADE distractors:</p> <ul style="list-style-type: none"> • I should buy these shoes because they will make me hug people • I should buy these shoes because I like to play your game <p>TRADE-control distractors:</p> <ul style="list-style-type: none"> • I should get an [car brand] because it is stylish. • I should consider [place name] for snack time because I can enjoy this with my boyfriend.
	<p>Text: Get a tasty look. By [brand name]</p> <p>Explanation: <i>I should go to [brand name] not only does their food taste great but it also looks good</i></p> <p>TRADE distractors:</p> <ul style="list-style-type: none"> • I should go to [brand name] because my eyelashes need a new look • I should go to [brand name] because tasty burgers must look like these eyelashes <p>TRADE-control distractors:</p> <ul style="list-style-type: none"> • I should head this Heart Research Centre message, because it alerts me that my body and its organs are the product of many environments and many lives • I should fund [healthcare system] because we should take back control

Figure 4: Examples from TRADE and TRADE-control, along with our transcription of the text (just for readability, not part of the dataset). Brands and logos are edited out in the paper examples for presentation purposes but are visible in TRADE.

judgements was 94%. The cases where both participants selected the target explanation were 270 (90%).

E Tested Models

Here we provide an overview of the models used in our experiments.

CLIP (Radford et al., 2021) is a contrastive model where image and text are separately encoded by two transformer-based models and then projected to the same vector space. CLIP is trained with a contrastive loss that minimizes the cosine distance between matching pairs of image and text embeddings. We used it in the Hugging Face implementation.¹¹

ALIGN (Jia et al., 2021) is also a contrastive vision-and-language model trained with the same loss function used for CLIP. It mainly differs from the latter in its encoders (EfficientNet for images and BERT for text) and in that also leverages noisy data during the training process. We used the Hugging Face model implementation.¹²

LiT (Zhai et al., 2022) is a contrastive model where the image encoder is “locked” (i.e. frozen) during pre-training, whereas the language encoder is initialized with random weights and trained from scratch with a contrastive loss. We used the Vision Transformer implementation¹³ by Google Research.

ALBEF (Li et al., 2021) is a vision-and-language model consisting of two separate transformer-based encoders from image and text and a multimodal encoder. The uni-modal modules are pre-trained contrastively and their outputs are then fused in the multimodal module, which is pre-trained with masked-language-modeling and image-text-matching objectives. We used the LAVIS implementation by Salesforce.¹⁴

¹¹https://huggingface.co/docs/transformers/model_doc/clip

¹²https://huggingface.co/docs/transformers/model_doc/align

¹³https://github.com/google-research/vision_transformer

¹⁴<https://github.com/salesforce/LAVIS>