A Limitations and societal impact

A.1 Limitations

BiVLC offers captions only in English. It would be interesting to extend the dataset to other languages, as some recent works in vision-language models are already doing [Pouget et al., 2024, Chen et al., 2023, Bugliarello et al., 2022]. Moreover, we only trained contrastive models, due to their suitability for image-to-text and text-to-image retrieval tasks and their availability. In the future, generative multimodal models, which we evaluated but not fine-tuned, could also be explored. Indeed, our approach to fine-tune contrastive models with hard negative images also has its limitations: we evaluated the models in Winoground [Thrush et al., 2022] and we saw that improvements are modest. There are several hypotheses to explain those results and one of them points to the effect of using synthetic images. Deeper analyses are needed to elucidate the real effect of synthetic images to train multimodal models. Finally, as we rely on SugarCrepe hard negative captions, we also use the same categories. Adding more diversity by extending BiVLC to other categories could be beneficial.

A.2 Societal impacts

Vision-language models such as CLIP are becoming popular models for many applications, but previous research has probed and analyzed their limitations [Yuksekgonul et al., 2022, Hsieh et al., 2024]. We contribute with our research by delving into a new point of view: the importance of measuring bidirectional compositionality. We hope that our benchmark will lead to a better assessment of the compositional understanding of vision-language models and may thus lead to their improvement. Furthermore, we expect BiVLC to become one of the main benchmarks used to improve the compositional capabilities of vision-language models, as it enables deeper analysis of their behaviour and it is more challenging than previous VLC tasks.

B BiVLC dataset information

We host BiVLC at HuggingFace[3]. The Croissant metadata record, which contains dataset metadata, is available in the dataset repository[4]. The DOI for BiVLC dataset is 10.57967/hf/2391, can be found in the dataset repository[5]. We provide a summary below.

Dataset documentation BiVLC is a benchmark for Bidirectional Vision-Language Compositionality evaluation. Each instance consists of two images and two captions. Using each of the images and captions as a base, a model is asked to select the pair that correctly represents the base versus the hard negative distractor with minor compositional changes. Thus, we can measure image-to-text and text-to-image retrieval with hard negative pairs. To obtain good results on the dataset, it is necessary that the model performs well in both directions for the same instance. Each instance of the dataset consists of six fields:

- image: COCO 2017 validation image.
- caption: COCO 2017 validation text describing the COCO image.
- negative_caption: Negative caption generated from the COCO text description by SugarCrepe [Hsieh et al., 2024].
- negative_image: Negative image generated from the negative caption by us for BiVLC.
- type: Category of the negative instances: Replace, Swap or Add.
- subtype: Subcategory of the negative instances: Object, Attribute or Relation.

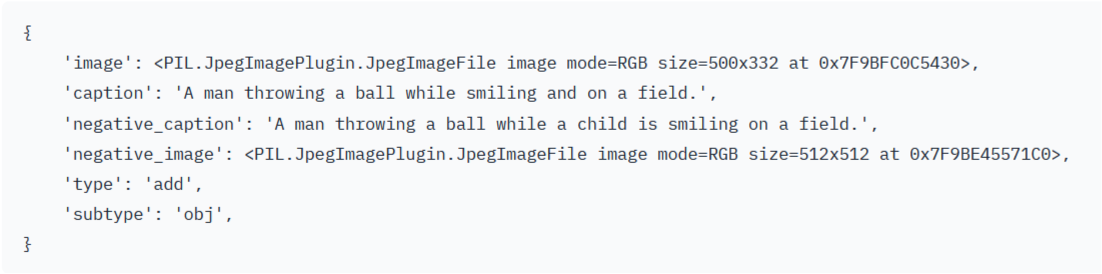
Example of a BiVLC instance after load_dataset("imirandam/BiVLC", split = "test") (Figure 5):

Maintenance plan We are committed to maintaining the dataset to resolve any technical issues. We actively track issues in the HuggingFace or GitHub repositories provided in `https://imirandam.github.io/BiVLC_project_page`.

[3] `https://huggingface.co/datasets/imirandam/BiVLC`

[4] `https://huggingface.co/api/datasets/imirandam/BiVLC/croissant`

[5] `https://huggingface.co/datasets/imirandam/BiVLC?doi=true`



```
{
    'image': <PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x332 at 0x7F9BFC0C5430>,
    'caption': 'A man throwing a ball while smiling and on a field.',
    'negative_caption': 'A man throwing a ball while a child is smiling on a field.',
    'negative_image': <PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=512x512 at 0x7F9BE45571C0>,
    'type': 'add',
    'subtype': 'obj',
}
```

Figure 5: Example of a BIVLC instance after loading the dataset.

Licensing We license our work using the MIT License [6].

Author statement We, the authors, assume full responsibility in case of violation of rights.

C TROHN-TEXT dataset details

The detailed statistics for the TROHN-TEXT dataset can be found in Table 6. The number of instances per category and subcategory are provided.

Table 6: Statistics for the TROHN-TEXT dataset, divided into the different categories and subcategories. Each instance is composed by one image and two captions.

	REPLACE			SWAP		ADD		TOTAL
	OBJ	ATT	REL	OBJ	ATT	OBJ	ATT	
# instances	570,325	571,609	576,101	333,705	429,163	584,077	587,866	3,652,846

D Detailed evaluation metrics

The **I2T** score measures the performance for image-to-text retrieval. For each instance in our dataset, we actually have two image-to-text retrieval examples. To obtain a perfect I2T score, the correct captions for both images have to be selected. Thus, assuming C_0, C_1 refer to positive and negative caption respectively, I_0, I_1 to positive and negative image, and we use $s(C_i, I_i)$ as the similarity function for a caption and an image, I2T score $I2T(C_0, I_0, C_1, I_1)$ is defined in Equation 1:

$$I2T\left(C_0, I_0, C_1, I_1\right) = \begin{cases} 1 & \text{if } s\left(C_0, I_0\right) > s\left(C_1, I_0\right) \\ & \text{and } s\left(C_1, I_1\right) > s\left(C_0, I_1\right) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The **T2I** score $T2I(C_0, I_0, C_1, I_1)$ is similarly defined for text-to-image retrieval (Equation 2):

$$T2I\left(C_0, I_0, C_1, I_1\right) = \begin{cases} 1 & \text{if } s\left(C_0, I_0\right) > s\left(C_0, I_1\right) \\ & \text{and } s\left(C_1, I_1\right) > s\left(C_1, I_0\right) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Finally, the **Group** score $G(C_0, I_0, C_1, I_1)$ is the main metric, since it combines the performance for image-to-text and text-to-image retrieval. To obtain a perfect group score for a given instance, both images have to be matched with the suitable captions and both captions with the suitable images. The group score is defined in Equation 3:

$$G\left(C_0, I_0, C_1, I_1\right) = \begin{cases} 1 & \text{if } I2T\left(C_0, I_0, C_1, I_1\right) \\ & \text{and } T2I\left(C_0, I_0, C_1, I_1\right) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

[6] `https://github.com/IMirandaM/BiVLC/blob/main/LICENSE`

E Implementation details

This appendix contains all the information related to the implementation of the experiments. All the source code with instructions can be found at `https://github.com/IMirandaM/BiVLC`.

E.1 Source datasets

We obtain all source datasets directly from the original sources published by the authors. To the best of our knowledge, all data sources we use are open to non-commercial use, do not contain personally identifiable information and do not contain offensive content.

- **COCO** [Lin et al., 2014]: We obtain COCO 2017 from the official project website[7] under a Creative Commons Attribution 4.0 License[8].
- **SUGARCREPE** [Hsieh et al., 2024]: We obtain the negative captions from the official GitHub project[9] under the MIT license[10].

E.2 Train and Validation datasets

Three different variants, $\text{CLIP}_{\text{COCO}}$, $\text{CLIP}_{\text{TROHN-TEXT}}$ and $\text{CLIP}_{\text{TROHN-IMG}}$, have been fine-tuned. For that we used 3 different training and validation datasets:

- $\text{CLIP}_{\text{COCO}}$ is fine-tuned using COCO 2017 train which contains 591,753 captions and 118,287 images, i.e., 591,753 instances formed by an image and a caption.
- $\text{CLIP}_{\text{TROHN-TEXT}}$ is fine-tuned with the TROHN-TEXT dataset consisting of 3,652,846 instances formed by one image and two captions. See Section 5.1 for more details.
- $\text{CLIP}_{\text{TROHN-IMG}}$ is fine-tuned with the TROHN-IMG dataset consisting of 296,070 instances formed by two images and two captions, i.e. 592,140 pairs, an amount similar to that of the COCO 2017 train. See Section 5.2 for more details.

All three datasets are randomly divided into 80% for training and 20% for validation. The TROHN-TEXT and the TROHN-IMG datasets are in HuggingFace repositories[11][12], with all the necessary information for its use.

E.3 Software information

Models We detail the sources of the pretrained and fine-tuned models we used.

- **OPENCHAT-3.5-0106** We obtain the model released by [Wang et al., 2023a] [13].
- **COLA** We obtain RoBERTa base model [Liu et al., 2019] fine-tuned in CoLA released by [Morris et al., 2020] [14].
- **VERA** We obtain pretrained Vera model released by [Liu et al., 2023][15].
- **Pretrained CLIP from OpenCLIP** We obtain the pretrained baseline VIT-B-32 OpenAI's CLIP model [Radford et al., 2021] from OpenCLIP [Cherti et al., 2022][16].
- Fine-tuned CLIP models:
    1. **NEGCLIP** We obtain fine-tuned CLIP model released by [Yuksekgonul et al., 2022][17].

[7] `https://cocodataset.org/#download`

[8] `https://cocodataset.org/#termsofuse`

[9] `https://github.com/RAIVNLab/sugar-crepe/tree/main/data`

[10] `https://github.com/RAIVNLab/sugar-crepe/blob/main/LICENSE`

[11] `https://huggingface.co/datasets/imirandam/TROHN-Text`

[12] `https://huggingface.co/datasets/imirandam/TROHN-Img`

[13] `https://huggingface.co/openchat/openchat-3.5-0106`

[14] `https://huggingface.co/textattack/roberta-base-CoLA`

[15] `https://huggingface.co/liujch1998/vera`

[16] `https://github.com/mlfoundations/open_clip`

[17] `https://github.com/mertyg/vision-language-models-are-bows`

2. **GNM** We obtain fine-tuned CLIP model released by [Sahin et al., 2024][18].

- **VQAScore** We obtain model released by [Lin et al., 2024][19].
- **Open CapPa** We obtain the open source CapPa model from the official reposiroty [20].
- **CapPa** SugarCrepe results obtained from [Tschannen et al., 2024].
- **GPT-4V** results are taken from the SugarCrepe web page [21].

Fine-tuning hyperparameters We fine-tuned three CLIP models: CLIP$_{\text{COCO}}$, CLIP$_{\text{TROHN-Text}}$ and CLIP$_{\text{TROHN-Img}}$. We also trained two detectors, CLIP$_{\text{Det}}$ and CLIP$_{\text{TROHN-Img/Det}}$, based on synthetic-natural image-text detection.

We base CLIP fine-tuning on OpenCLIP [Cherti et al., 2022]. Detailed hyperparameters:

- Learning rate: 1e-6.
- Scheduler: Cosine scheduler with 50 warmup steps.
- Optimizer: AdamW optimizer with beta1 = 0.9, beta2 = 0.98, eps = 1e-6 and weight decay = 0.1.
- Loss function: InfoNCE Loss. In the case of CLIP$_{\text{TROHN-Text}}$ the loss is modified to add only negative captions following the idea proposed in NegCLIP [Yuksekgonul et al., 2022].
- Batch size: We define a batch size of 400 (400 images x 400 captions) for CLIP$_{\text{COCO}}$. For CLIP$_{\text{TROHN-Text}}$ and CLIP$_{\text{TROHN-Img}}$ we define a batch size of 200, and then we add negatives. In the case of CLIP$_{\text{TROHN-Text}}$, as it has not hard negative images, it results in 200 images x 400 captions (positive + hard negatives). For CLIP$_{\text{TROHN-Img}}$ the batch consists of 200 positive pairs and 200 negative pairs, resulting in 400 images x 400 captions.
- Epochs: We fine-tune all models over 10 epochs and we used validation accuracy as the model selection criterion, i.e. we selected the model with the highest accuracy on the corresponding validation set.

We also trained CLIP$_{\text{Det}}$ and CLIP$_{\text{TROHN-Img/Det}}$ detectors for binary classification by keeping the encoders frozen and adding a sigmoid neuron over the CLS embedding for the image encoder and over the EOT embedding for the text encoder. Detailed hyperparameters:

- Learning rate: 1e-6.
- Optimizer: Adam optimizer with beta1 = 0.9, beta2 = 0.999, eps = 1e-08 and without weight decay.
- Loss function: Binary cross-entropy loss (BCELoss).
- Batch size: We define a batch size of 400.
- Epochs: We trained the text detector over 10 epochs and the image detectors over 1 epoch. We used validation accuracy as the model selection criterion, i.e. we selected the model with the highest accuracy in the corresponding validation set.

Evaluation For contrastive models, we base our evaluation on OpenCLIP [Cherti et al., 2022]. We follow all the default hyperparameters used to evaluate models, making sure that when loading the checkpoints we are using QuickGELU as in the base pretrained model[22]. For the generative models, we have used the official GitHub repositories of each model and followed the instructions and relied on the evaluation codes provided by their authors.

[18] `https://github.com/ugorsahin/Generative-Negative-Mining`

[19] `https://github.com/linzhiqiu/t2v_metrics`

[20] `https://github.com/borisdayma/clip-jax`

[21] `https://github.com/RAIVNLab/sugar-crepe/tree/main/gpt-4v-results`

[22] Evaluation bug when using GELU vs QuickGELU `https://github.com/RAIVNLab/sugar-crepe/issues/7`

E.4 Hardware information

All the experiments of this research have been performed on internal clusters of HiTZ Zentroa. Experiments are run on two servers. For the main experiments, we work in a Supermicro Superserver AS-4124GO-NART with 2 x AMD EPYC 7513, 1024 GB RAM and 8x A100-SXM4-80GB GPUs. For smaller runs, we work in a Dell Poweredge R740 with 2x Intel Xeon Gold 6226R, 256 GB RAM and 2x A30 GPUs. We estimate that for the generation of the datasets (BIVLC, TROHN-TEXT and TROHN-IMG), and the training of the different models, around 387 hours of execution have been needed. Detailed information for specific runs:

CLIP fine-tuning For fine-tuning CLIP models we use one NVIDIA A100-SXM4-80GB GPU. The execution time varies depending on the variant: for $\text{CLIP}_{\text{TROHN-TEXT}}$, due to the amount of data, it takes about 6 days, in the case of $\text{CLIP}_{\text{COCO}}$ and $\text{CLIP}_{\text{TROHN-IMG}}$ as the datasets are smaller about 18 hours each.

BIVLC It was necessary to create 30,048 images in total. For that purpose, we used 4 NVIDIA A100-SXM4-80GB GPUs with an approximate execution time of 11 hours.

TROHN-TEXT dataset We used one NVIDIA A100-SXM4-80GB GPU. For the generation of each of the seven subtypes proposed in SUGARCREPE it takes approximately 16 hours, estimating a total of 112 hours.

TROHN-IMG dataset It was necessary to generate 296,070 images. Thus we used 6 NVIDIA A100-SXM4-80GB GPUs with an execution time of approximately 72 hours.

Detectors We used one NVIDIA A30 GPU. The sigmoid neuron of the text detector is trained for 10 epochs taking approximately 1.5 hours, and for the image detector it is trained for one epoch taking approximately 3 hours.

Evaluation All evaluations have been performed on one NVIDIA A100-SXM4-80GB GPU.

F SUGARCREPE templates

For the generation of hard negative captions, for the training and validation splits, we used the templates proposed by SUGARCREPE. Each subcategory defined in SUGARCREPE has a template that can be found in [Hsieh et al., 2024].

G BIVLC instance examples

Examples of each category and their corresponding subcategories are presented in Figure 6. If you wish to see more examples in a clearer way, please access the viewer mode of the BiVLC repository[23].

[23] https://huggingface.co/datasets/imirandam/BiVLC/viewer

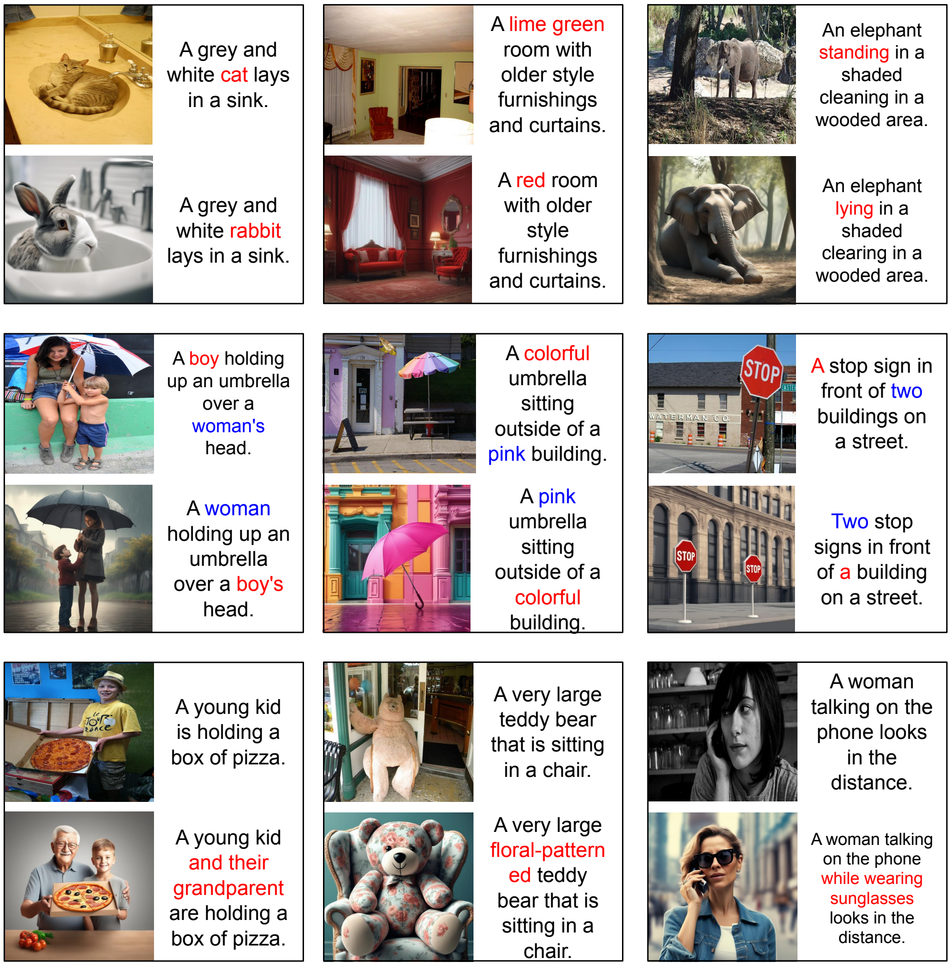

A grey and white cat lays in a sink.
A grey and white rabbit lays in a sink.
A lime green room with older style furnishings and curtains.
A red room with older style furnishings and curtains.
An elephant standing in a shaded cleaning in a wooded area.
An elephant lying in a shaded clearing in a wooded area.
A boy holding up an umbrella over a woman's head.
A woman holding up an umbrella over a boy's head.
A colorful umbrella sitting outside of a pink building.
A pink umbrella sitting outside of a colorful building.
A stop sign in front of two buildings on a street.
Two stop signs in front of a building on a street.
A young kid is holding a box of pizza.
A young kid and their grandparent are holding a box of pizza.
A very large teddy bear that is sitting in a chair.
A very large floral-patterned teddy bear that is sitting in a chair.
A woman talking on the phone looks in the distance.
A woman talking on the phone while wearing sunglasses looks in the distance.

Figure 6: BiVLC instance examples. From top to bottom three examples of the categories Replace (first row), Swap (second row) and Add (third row).

H Crowdsourcing

We have carried out crowdsourcing on the Prolific platform[24]. The following sections will detail the instructions and payments.

H.1 Instructions for the different stages

We have used crowdsourcing in three different stages, here are the instructions for each stage.

Choose the best generated image The instructions for this stage are detailed below (see Step 3 in Section 3). An example of a question as seen by the annotators is depicted in Figure 7.

"Each instance will consist of a text description and 4 images (A,B,C,D). The aim is to choose the image that represents the information provided in the description, and discard the rest. It might be that more than one image is acceptable or none is acceptable. We only need you to select one, if available, or none.

[24] https://www.prolific.com/

Instructions:

Select one image that adequately represents the description. In case none of the images adequately represent the description, select the answer "None".

Important:

1) All information presented in the text description must appear in the image.

2) Aesthetic defects are accepted, such as deformed faces, arms, hands etc."

Filter ambiguous instances The instructions for this stage are detailed below (see Step 4 in Section 3). An example of a question as seen by the annotators is depicted in Figure 8.

"Each instance will consist of a description and two images. The objective is to choose the image that best represents the description or, in case both images represent the description equally well, choose "Both".

Instructions:

Choose the image that best matches the description, in the case that both images match the description equally well choose "Both".

Important:

1) Aesthetic defects are accepted, such as deformed faces, arms, hands etc."

Human Baseline The instructions for this stage are divided into two: image-to-text retrieval (I2T) and text-to-image retrieval (T2I). Both are detailed below, and examples of question as seen by the annotators are depicted in Figures 9 and 10. The obtained human performance for each task can be seen in Table 2.

"Each survey consists of 100 questions/instances. We split the instances into 2 parts (50/50), in which the task changes.

The first part of the instances consists of an image and two descriptions. The goal is to choose the description that BEST represents the image.

1) Instructions: Choose the description that BEST represents the image.
2) Important: Pay close attention to word order.

The second part of the instances consists of a description and two images. The objective is to choose the image that BEST represents the description.

1) Instructions: Choose the image that BEST matches the description.
2) Important: Aesthetic defects are accepted, such as deformed faces, arms, hands etc."

H.2 Hourly wage paid to participants and total amount spent

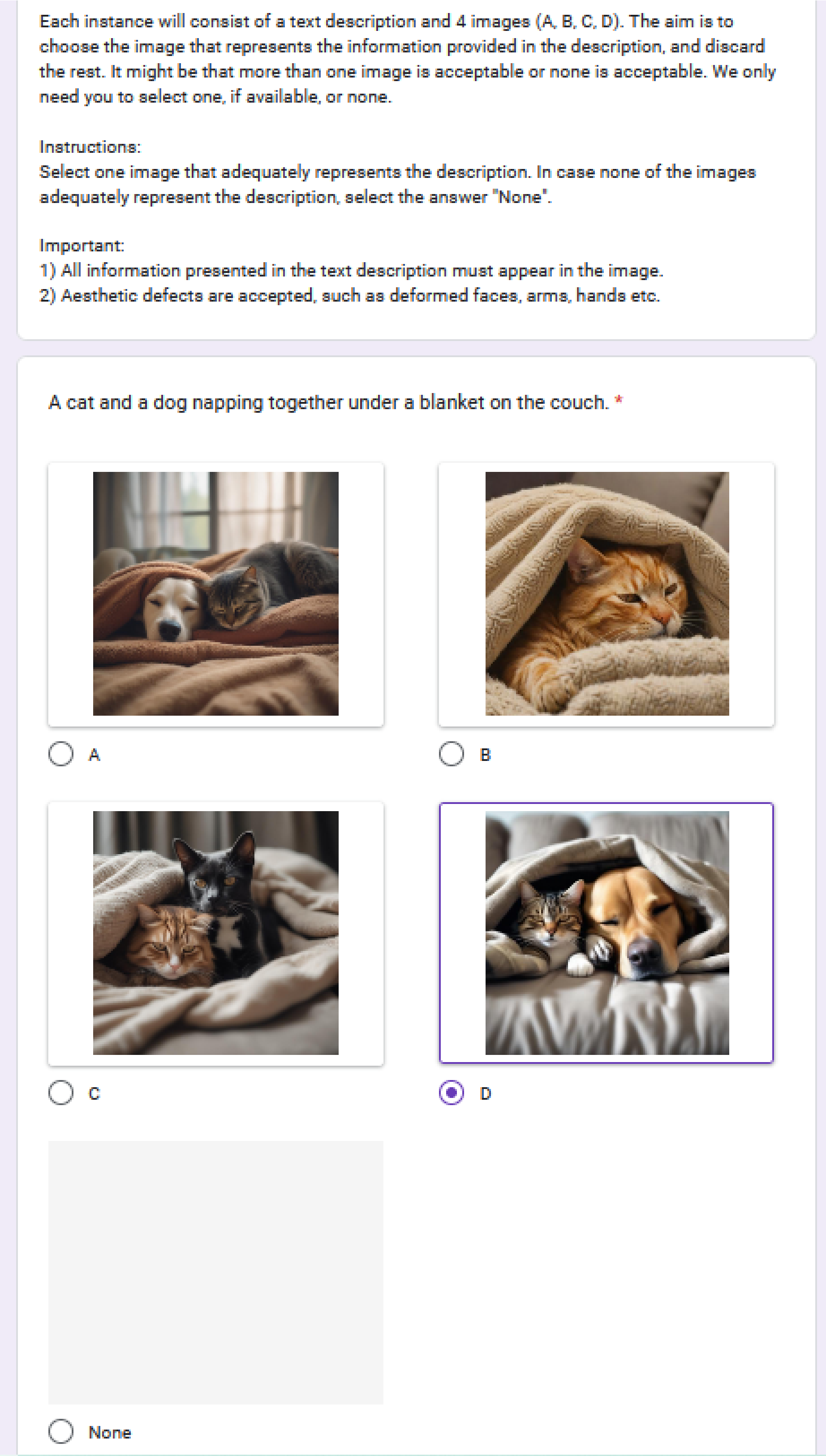
The hourly wage paid to all participants has been US$12.0, the recommended rate in the Prolific platform. The total amount spent on crowdsourcing is US$1,331.61, which includes the payment to participants plus service commissions.

H.3 Inter-tagger agreement

During "Step 3 - Ask to human annotators to choose the best generated image" (see Section 3.1) we used 10 annotators divided into pairs to score a total of 500 annotations and obtain inter-tagger agreement. To do this, we calculated Cohen's Kappa score between the responses of each pair who took the same survey (Table 7). The score is calculated based on whether they chose one of the images or none. The mean score obtained among the 5 groups of annotators is 0.49, which was interpreted as moderate agreement.

Table 7: Cohen's Kappa score for each of the pairs used to calculate inter-tagger agreement and the mean score of the 5 pairs.

Pair	Kappa score	Mean
1	0.66	**0.49**
2	0.43	
3	0.56	
4	0.43	
5	0.38	



Each instance will consist of a text description and 4 images (A, B, C, D). The aim is to choose the image that represents the information provided in the description, and discard the rest. It might be that more than one image is acceptable or none is acceptable. We only need you to select one, if available, or none.

Instructions:
Select one image that adequately represents the description. In case none of the images adequately represent the description, select the answer "None".

Important:
1) All information presented in the text description must appear in the image.
2) Aesthetic defects are accepted, such as deformed faces, arms, hands etc.

A cat and a dog napping together under a blanket on the couch. *

○ A

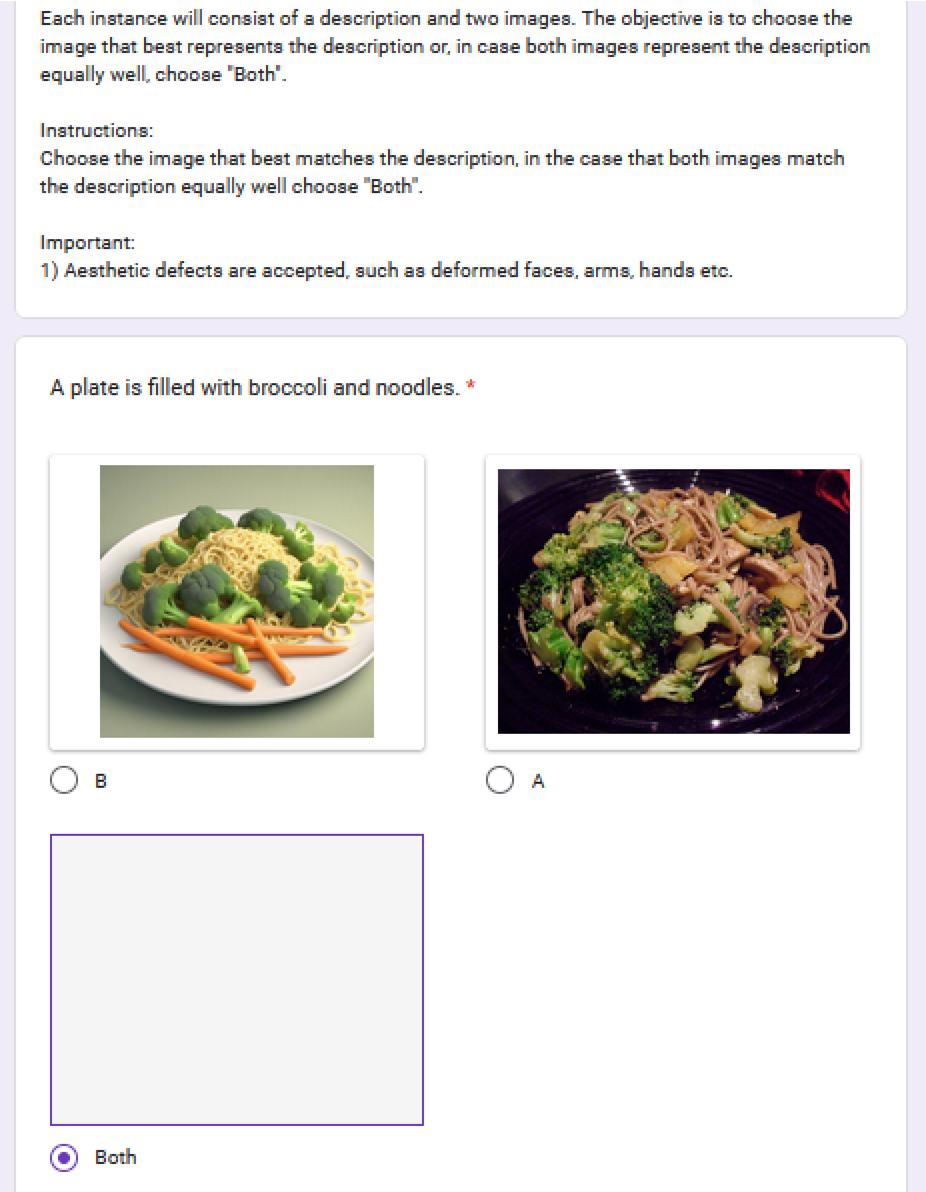
○ B

○ C

● D

○ None

Figure 7: Example provided in one of the surveys conducted to select the best negative image.



Each instance will consist of a description and two images. The objective is to choose the image that best represents the description or, in case both images represent the description equally well, choose "Both".

Instructions:
Choose the image that best matches the description, in the case that both images match the description equally well choose "Both".

Important:
1) Aesthetic defects are accepted, such as deformed faces, arms, hands etc.

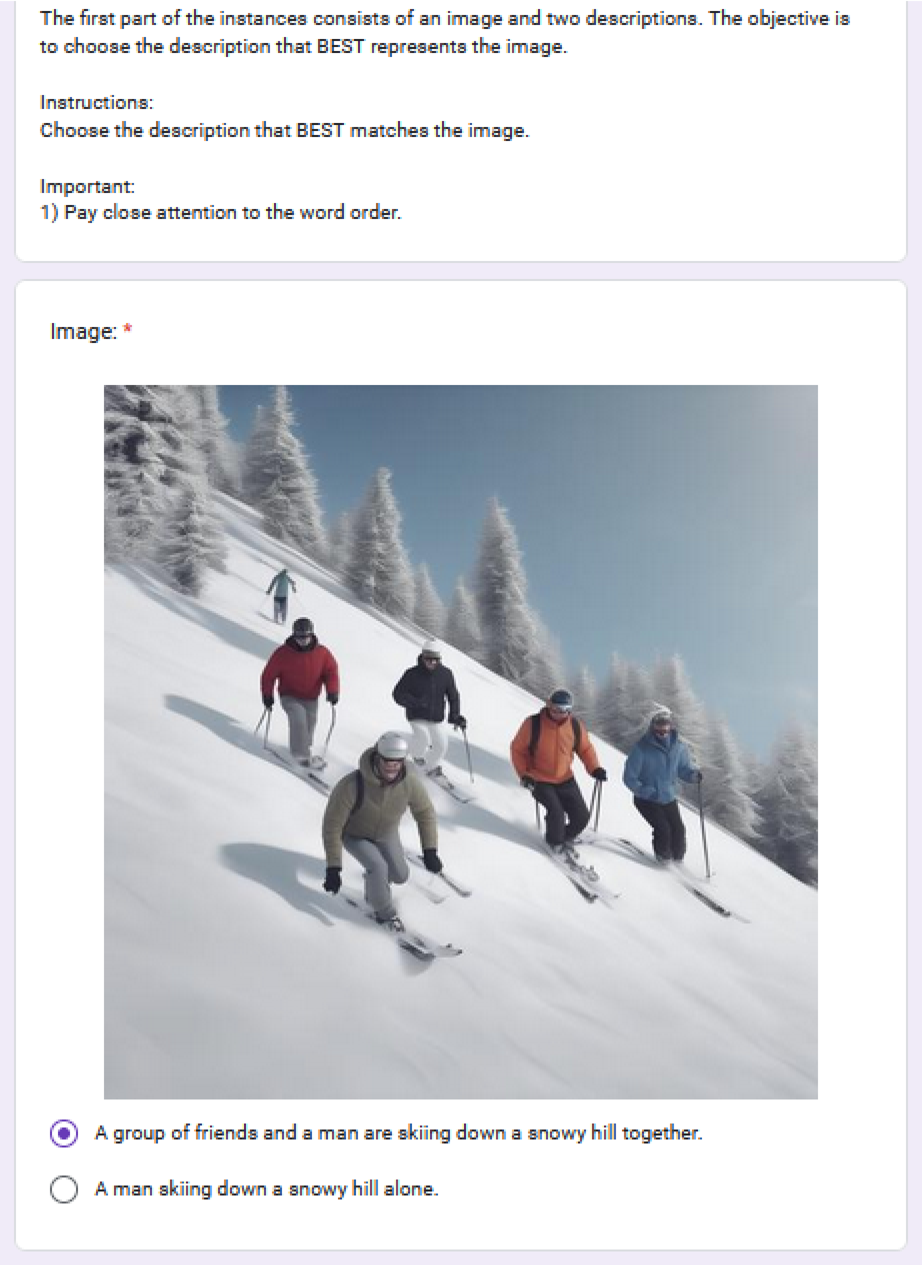
A plate is filled with broccoli and noodles. *

○ B

○ A

● Both

Figure 8: Example provided in one of the surveys conducted to disambiguate instances based on the original caption.



The first part of the instances consists of an image and two descriptions. The objective is to choose the description that BEST represents the image.

Instructions:
Choose the description that BEST matches the image.

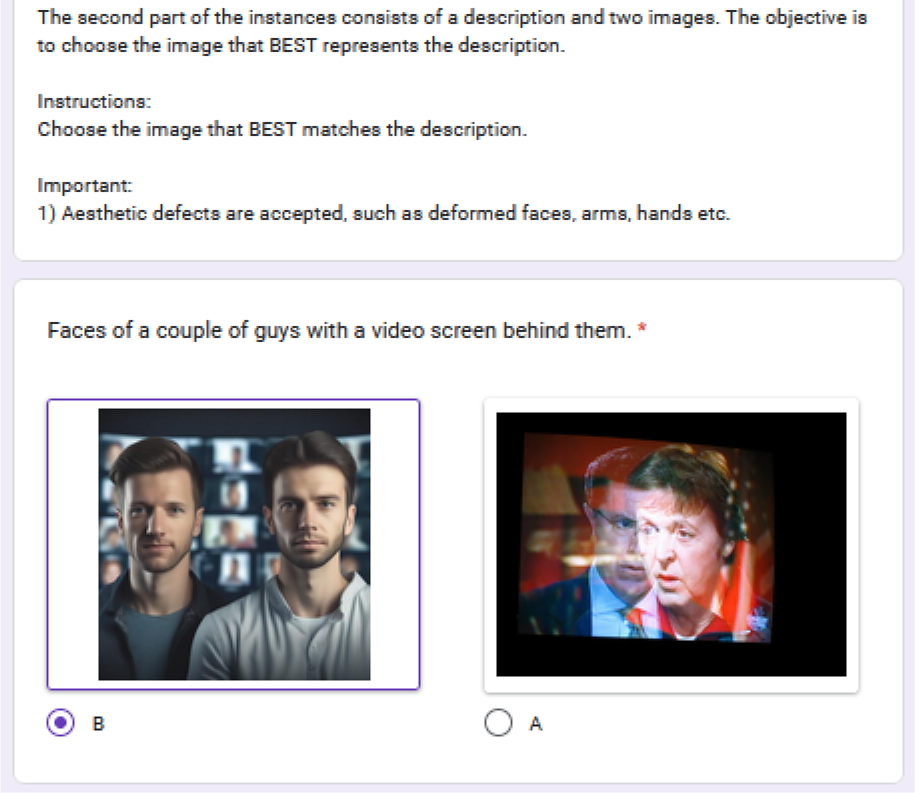
Important:
1) Pay close attention to the word order.

Image: *

● A group of friends and a man are skiing down a snowy hill together.

○ A man skiing down a snowy hill alone.

Figure 9: Example provided in one of the surveys conducted to obtain the human baseline in the I2T direction.



The second part of the instances consists of a description and two images. The objective is to choose the image that BEST represents the description.

Instructions:
Choose the image that BEST matches the description.

Important:
1) Aesthetic defects are accepted, such as deformed faces, arms, hands etc.

Faces of a couple of guys with a video screen behind them. *

● B

○ A

Figure 10: Example provided in one of the surveys conducted to obtain the human baseline in the T2I direction.