Generation of Captions Highlighting the Differences Between a Clothing Image Pair with Attribute Prediction

Kohei Abe[1], Soichiro Yokoyama[2], Tomohisa Yamashita[2], Hidenori Kawamura[2]

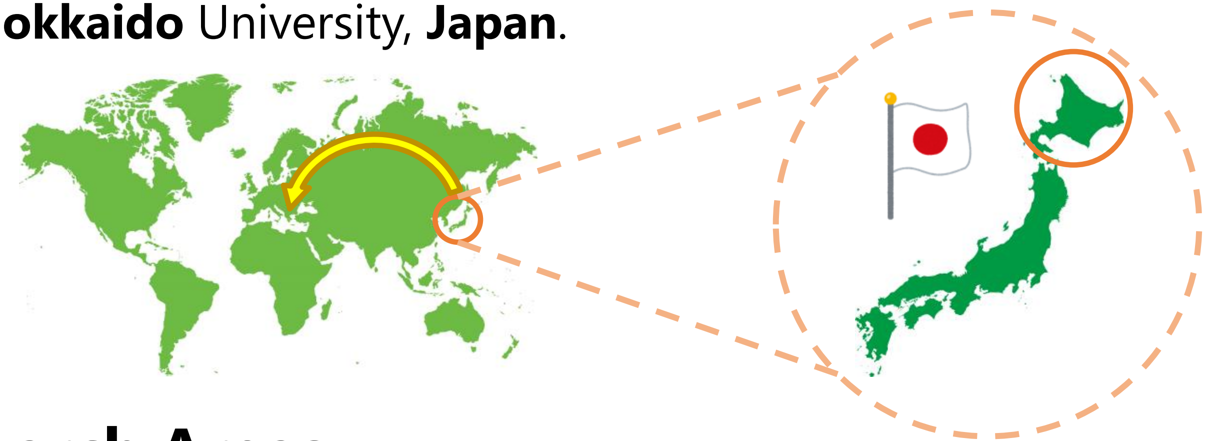
[1]Graduate School of Information Science and Technology, Hokkaido University, Japan
[2]Faculty of Information Science and Technology, Hokkaido University, Japan
Contact email: ko.abe@ist.hokudai.ac.jp

The Thirteenth International Conference on Intelligent Systems and Applications
INTELLI 2024
March 12, 2024, 12:30 - 14:00, Athens, Greece

Presenter: Kohei Abe

Education

- First year Master, Graduate School of Information Science and Technology, **Hokkaido** University, **Japan**.



Research Areas

- Image recognition, image caption generation, fashion image caption generation

Past Conference Participation & Publications

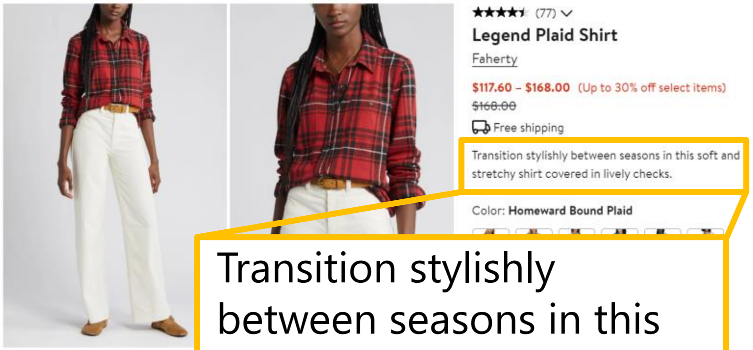
- “Generation of Captions Expressing the Bilateral Relationship between Pairs of Clothing Images using Attribute Estimation”, 22nd Forum on Information Technology (FIT), 2023.
- “Generation of Captions Expressing the Relationship between Pairs of Clothing Images”, Monthly Image Lab, 2023.

Background

- Consumer purchase process on e-commerce
  - Product comparison and evaluation
  - Detailed information is required.
  - Main sources of information
    - Product descriptions, user reviews, expert opinions, comparison websites, etc.

- **Product Descriptions**
  - Textual depiction of product features
  - Key source of information in the initial stage
  - Current situation
    - Single product focus
  - Problem
    - **Inadequate description of the differences between products**

      Lack of information to compare products



★★★★½ (77)

Legend Plaid Shirt

Faherty

$117.60 – $168.00 (Up to 30% off select items)

~~$168.00~~

Free shipping

Transition stylishly between seasons in this soft and stretchy shirt covered in lively checks.

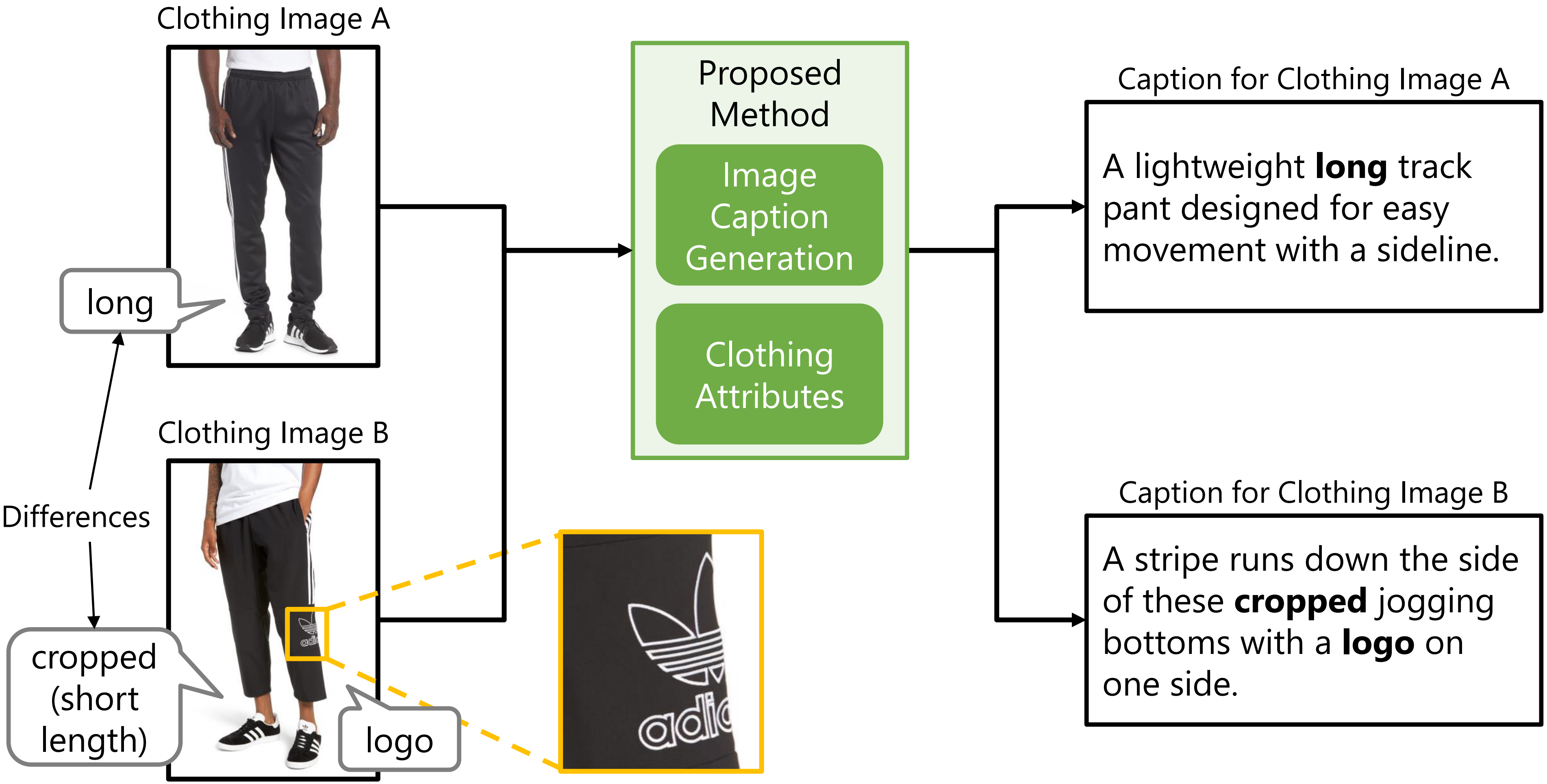
Color: Homeward Bound Plaid

Transition stylishly between seasons in this soft and stretchy shirt covered in lively checks.

An example of product descriptions
(Cited by Nordstrom)

Research Objective

- Development of a system to provide relevant information to consumers comparing clothing
    - **Generation of captions highlighting the differences between a clothing image pair**



Clothing Image A

long

Differences

Clothing Image B

cropped (short length)

logo

Proposed Method

Image Caption Generation

Clothing Attributes

Caption for Clothing Image A

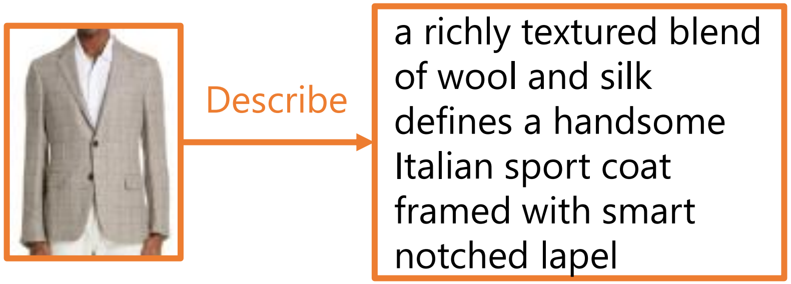
A lightweight **long** track pant designed for easy movement with a sideline.

Caption for Clothing Image B

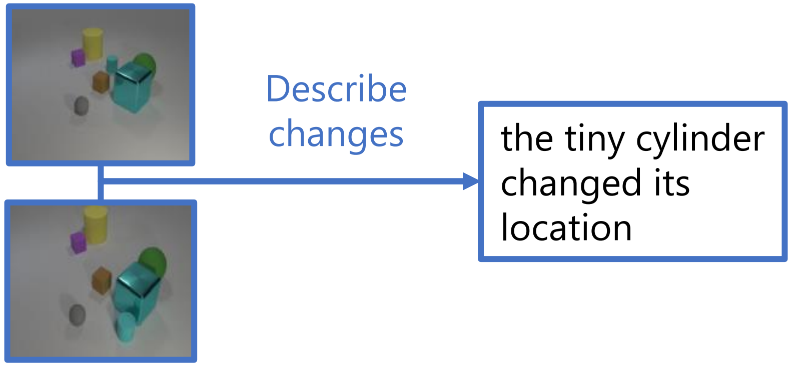
A stripe runs down the side of these **cropped** jogging bottoms with a **logo** on one side.

Related Work

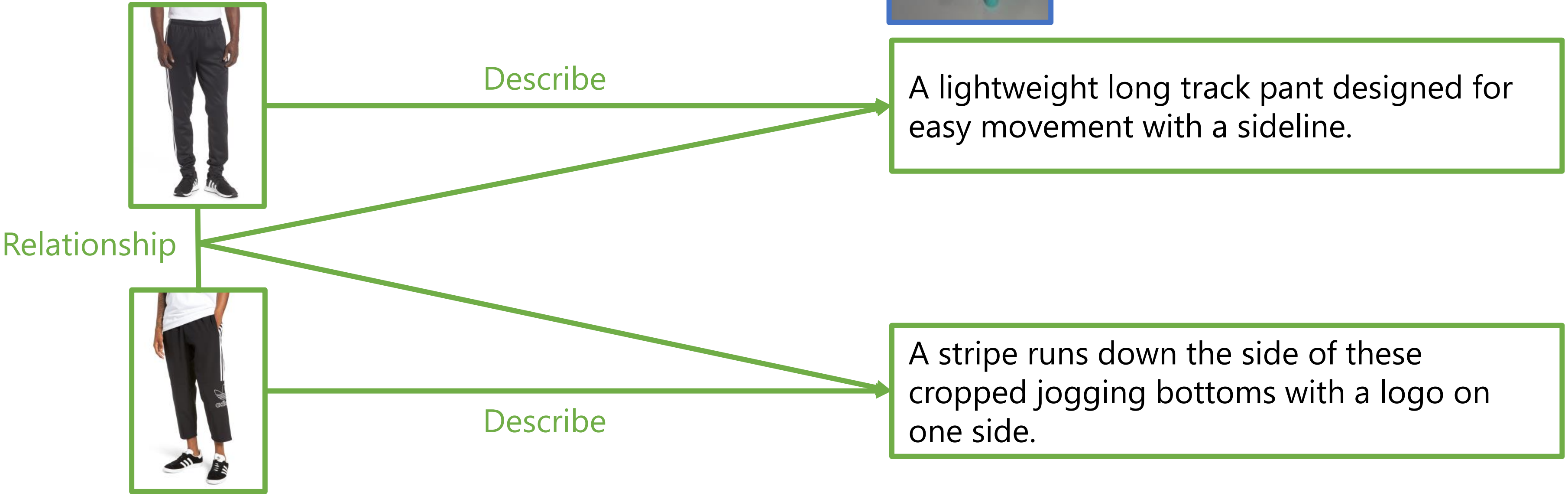
- Comparison of input-output relations between related and this research
  - Fashion Image Caption Generation[1]
    - Describing a clothing image
  - Change Image Caption Generation[2]
    - Describing changes between two images
  - **This Research**
    - Describing a clothing image, considering the relationship between the image pair


Describe

a richly textured blend of wool and silk defines a handsome Italian sport coat framed with smart notched lapel


Describe changes

the tiny cylinder changed its location


Describe

A lightweight long track pant designed for easy movement with a sideline.

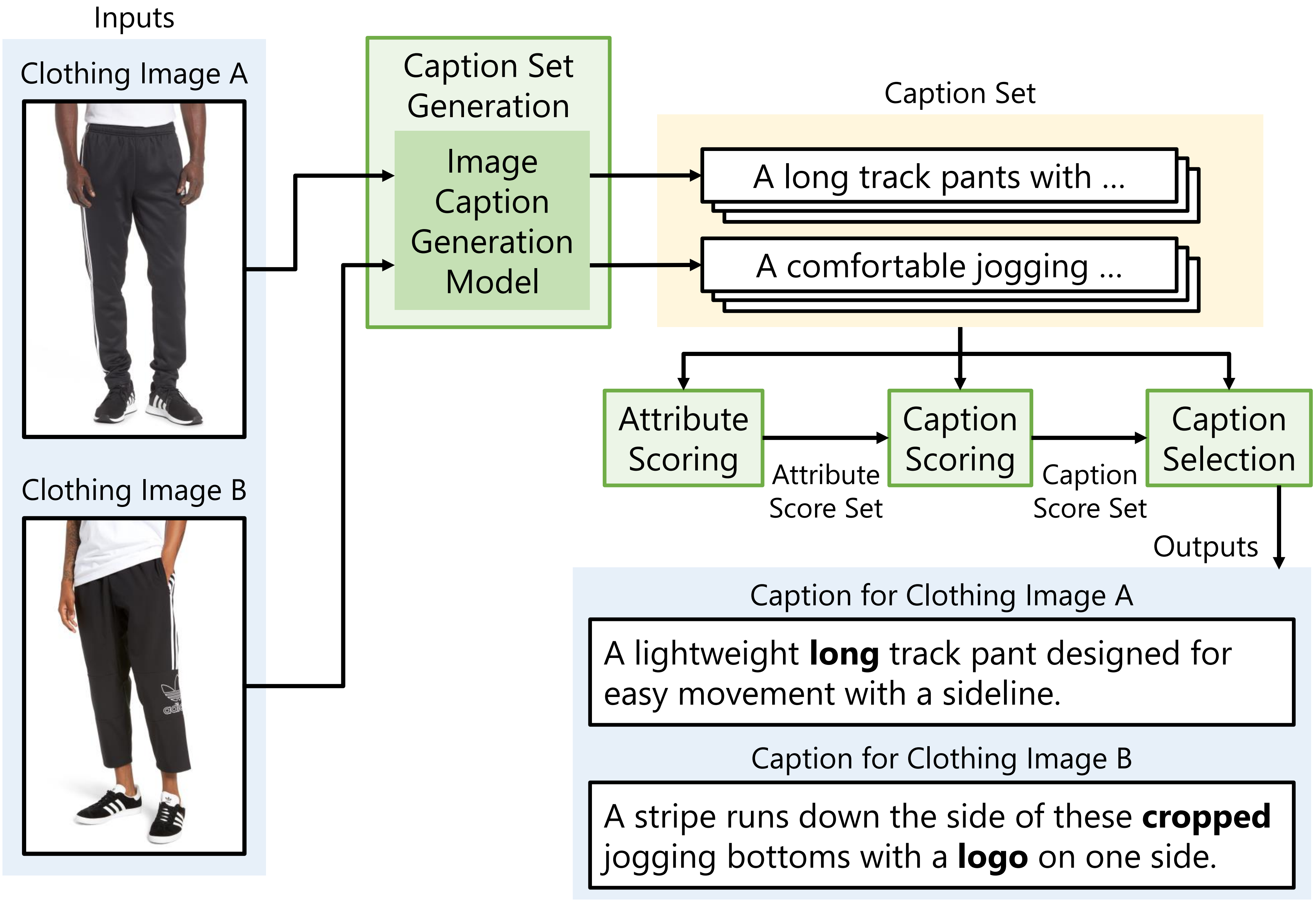
Relationship

Describe

A stripe runs down the side of these cropped jogging bottoms with a logo on one side.

[1] X. Yang et al., “Fashion captioning: Towards generating accurate descriptions with semantic rewards,” Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, Proceedings, Part XIII 16, pp. 1-17.
[2] D. H. Park, T. Darrell, and A. Rohrbach, “Robust change captioning,” Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4624-4633, 2019.

Overview of the Proposed Method



Inputs

Clothing Image A

Clothing Image B

Caption Set Generation

Image Caption Generation Model

Caption Set

A long track pants with …

A comfortable jogging …

Attribute Scoring

Attribute Score Set

Caption Scoring

Caption Score Set

Caption Selection

Outputs

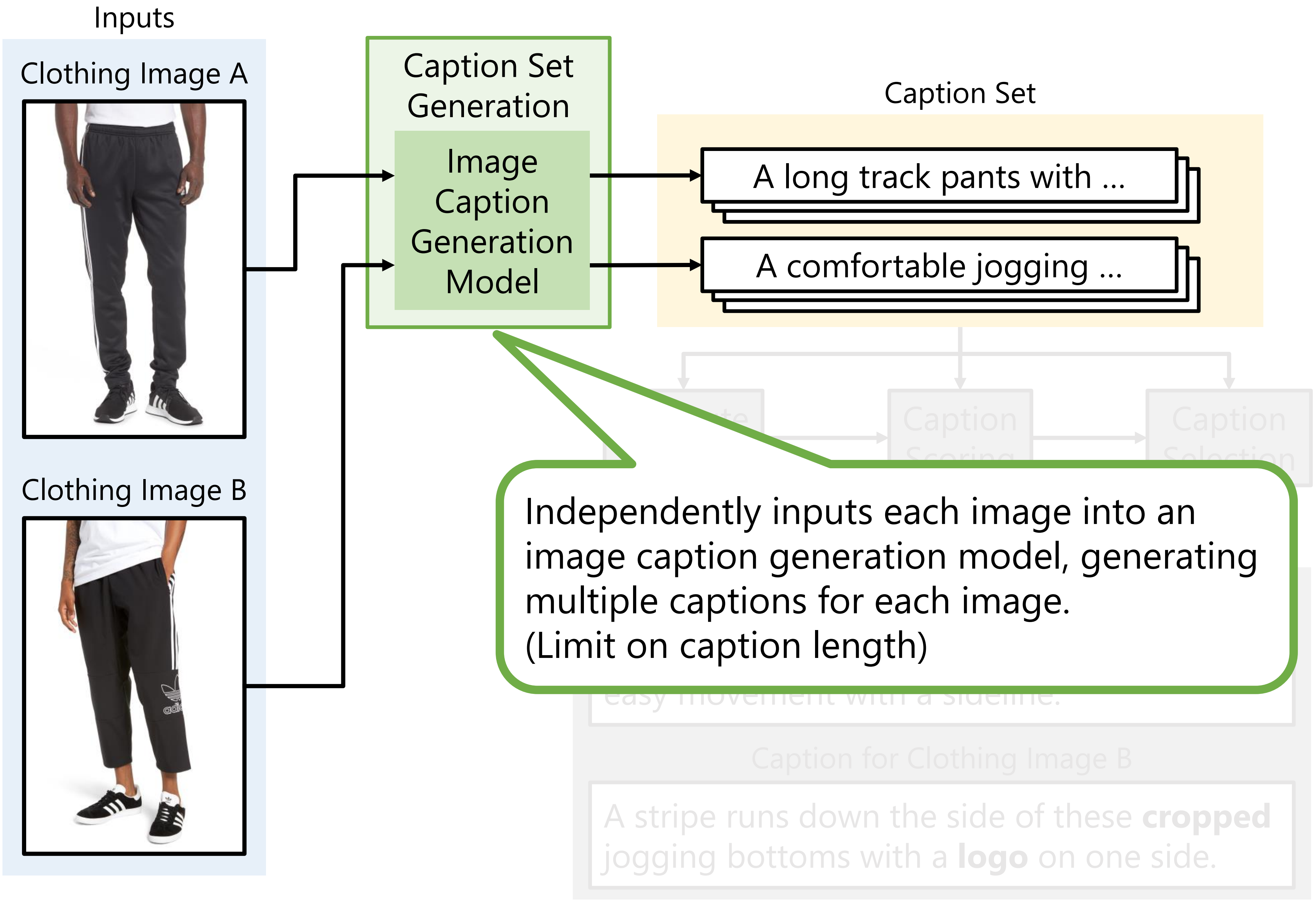
Caption for Clothing Image A

A lightweight **long** track pant designed for easy movement with a sideline.

Caption for Clothing Image B

A stripe runs down the side of these **cropped** jogging bottoms with a **logo** on one side.

Caption Set Generation Module



Independently inputs each image into an image caption generation model, generating multiple captions for each image.
(Limit on caption length)

easy movement with a sideline.

Caption for Clothing Image B

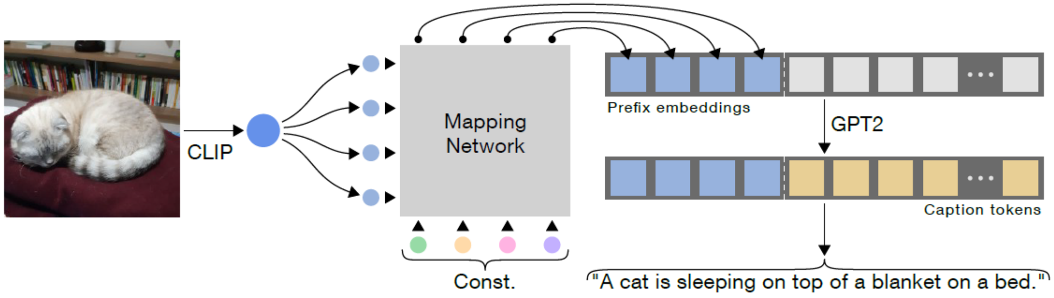
A stripe runs down the side of these **cropped** jogging bottoms with a **logo** on one side.

Image Caption Generation Model

Comparison of Image Caption Generation Models (MS COCO dataset)

Models	BLEU4	METEOR	Overview
NIC[3]	27.7	23.7	A combination of CNN and LSTM.
NICA[4]	25.0	23.9	Introduction of an attention mechanism to NIC.
SCST[5]	30.9	24.5	Use of reinforcement learning.
ClipCap[6]	**39.5**	**30.5**	**A combination of CLIP and GPT-2.**
OFA[7]	44.9	32.5	Pre-training using a large image-text pair.

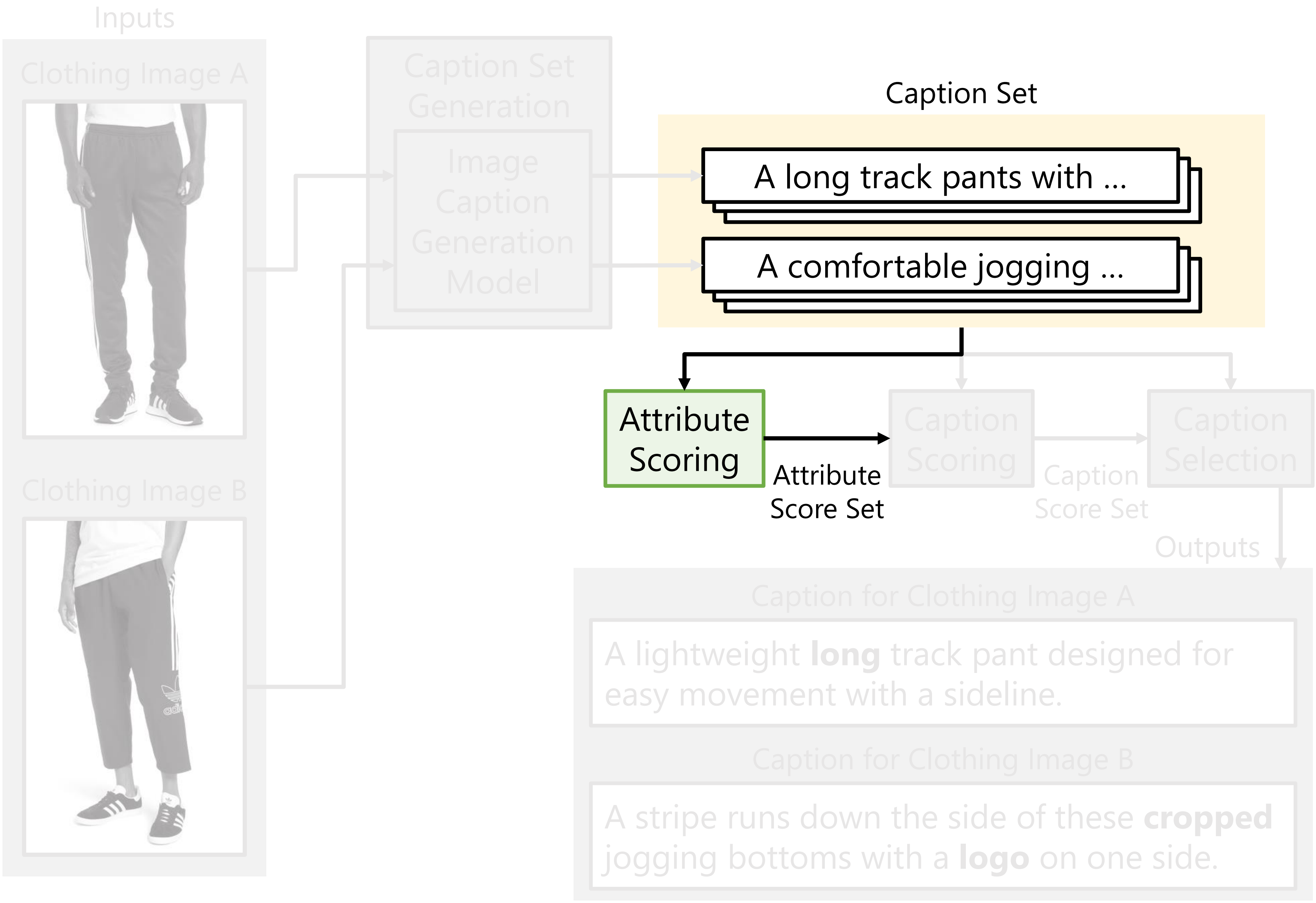
- Adopted **ClipCap**
  - Potential for multilingual support with multilingual CLIP[8] and GPT-4[9]
  - High performance among major models



Overview of ClipCap

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156-3164, 2015.
[4] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," International conference on machine learning, pp. 2048-2057, 2015.
[5] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Selfcritical sequence training for image captioning," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7008-7024, 2017.
[6] R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: Clip prefix for image captioning," arXiv preprint arXiv:2111.09734, 2021.
[7] P. Wang et al., "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," International Conference on Machine Learning, pp. 23318-23340, 2022.
[8] F. Carlsson, P. Eisen, F. Rekathati, and M. Sahlgren, "Cross-lingual and Multilingual CLIP," Proceedings of the Language Resources and Evaluation Conference, pp. 6848-6854, 2022.
[9] OpenAI, "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023.

Attribute Scoring Module



Inputs
Clothing Image A
Clothing Image B
Caption Set Generation
Image Caption Generation Model
Caption Set
A long track pants with …
A comfortable jogging …
Attribute Scoring
Attribute Score Set
Caption Scoring
Caption Score Set
Caption Selection
Outputs
Caption for Clothing Image A
A lightweight **long** track pant designed for easy movement with a sideline.
Caption for Clothing Image B
A stripe runs down the side of these **cropped** jogging bottoms with a **logo** on one side.

Attribute Scoring Module

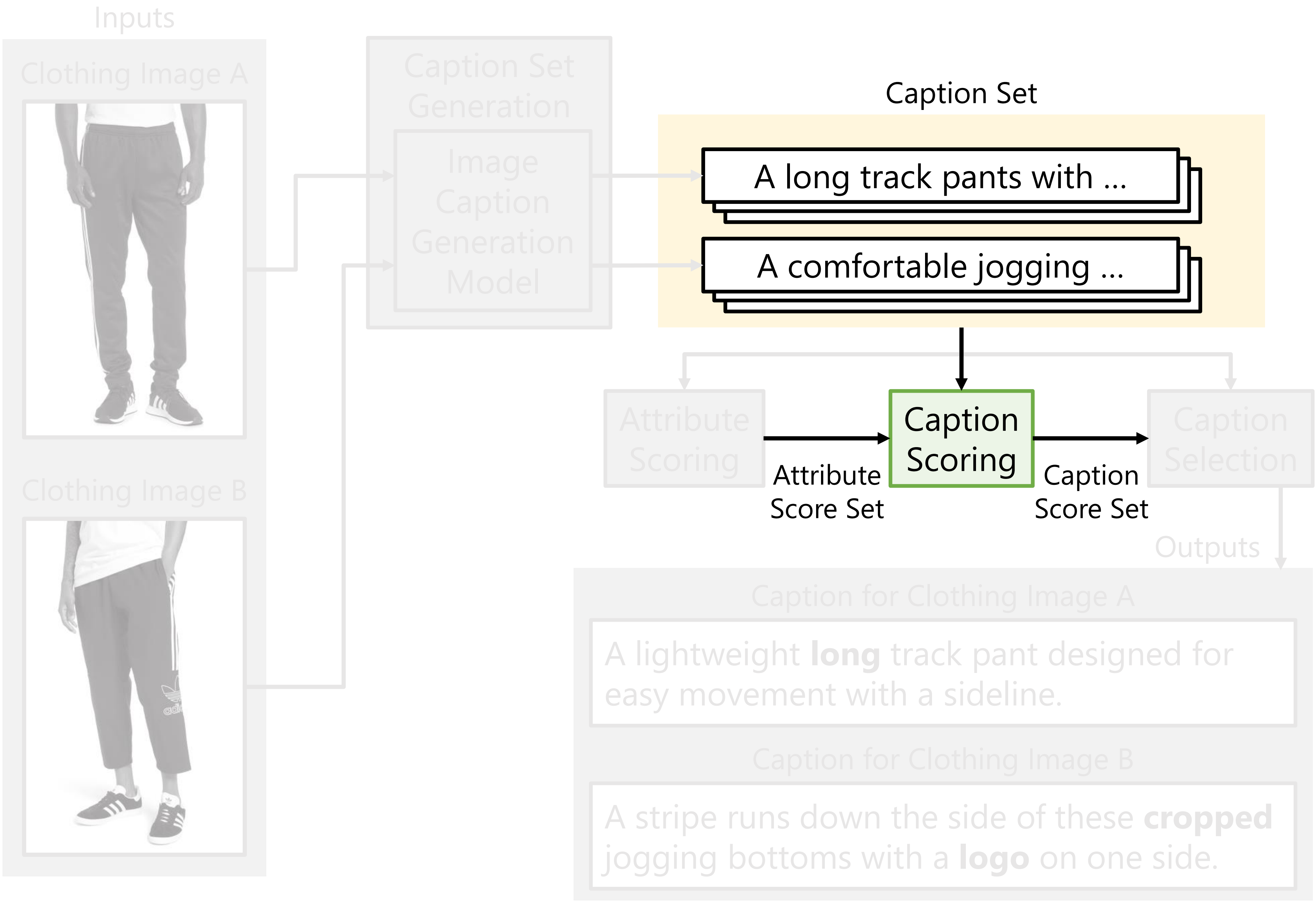
- Attributes[1]
  - Words that describe clothing features
  - Examples: cotton, leather, logo, sleeveless, stripe, graphic.
- Calculation of attribute scores for each clothing image
  - **Attribute Score**
    - Numerical expression of the prominence of a particular attribute exhibited by a clothing image
  - Using the frequency of occurrence of each attribute in the caption set for each image

Attribute Score Set

Attributes	Clothing Image	
	A	B
black	**0.9**	0.8
long	0.8	0.4
logo	0.3	0.8
cropped	0.2	0.7
⋮	⋮	⋮

The word "black" is contained in 90% of the captions for clothing image A.

Caption Scoring Module



Inputs

Clothing Image A

Clothing Image B

Caption Set Generation

Image Caption Generation Model

Caption Set

A long track pants with …

A comfortable jogging …

Attribute Scoring

Caption Scoring

Caption Selection

Attribute Score Set

Caption Score Set

Outputs

Caption for Clothing Image A

A lightweight **long** track pant designed for easy movement with a sideline.

Caption for Clothing Image B

A stripe runs down the side of these **cropped** jogging bottoms with a **logo** on one side.

Caption Scoring Module

- Calculation of caption score for each caption in caption set
  - **Caption Score**
    - Numerical expression of the extent to which the caption reflects the attribute differences between the clothing images and the attributes specific to each image.

① Calculate the **Relative Attribute Score** by subtracting one clothing image's attribute score from the another's

② Sum the relative attribute scores for the attributes included in each caption

Attribute Score Set

Attributes	Clothing Image	
	A	B
black	0.9	0.8
long	0.8	0.4
logo	0.3	0.8
cropped	0.2	0.7
⋮	⋮	⋮

①

Relative Attribute Score Set

Clothing Image	
A	B
0.1	-0.1
0.4	-0.4
-0.5	0.5
-0.5	0.5
⋮	⋮

□ − □

□ − □

Caption for clothing image A

A **long** track pants with a **logo** …

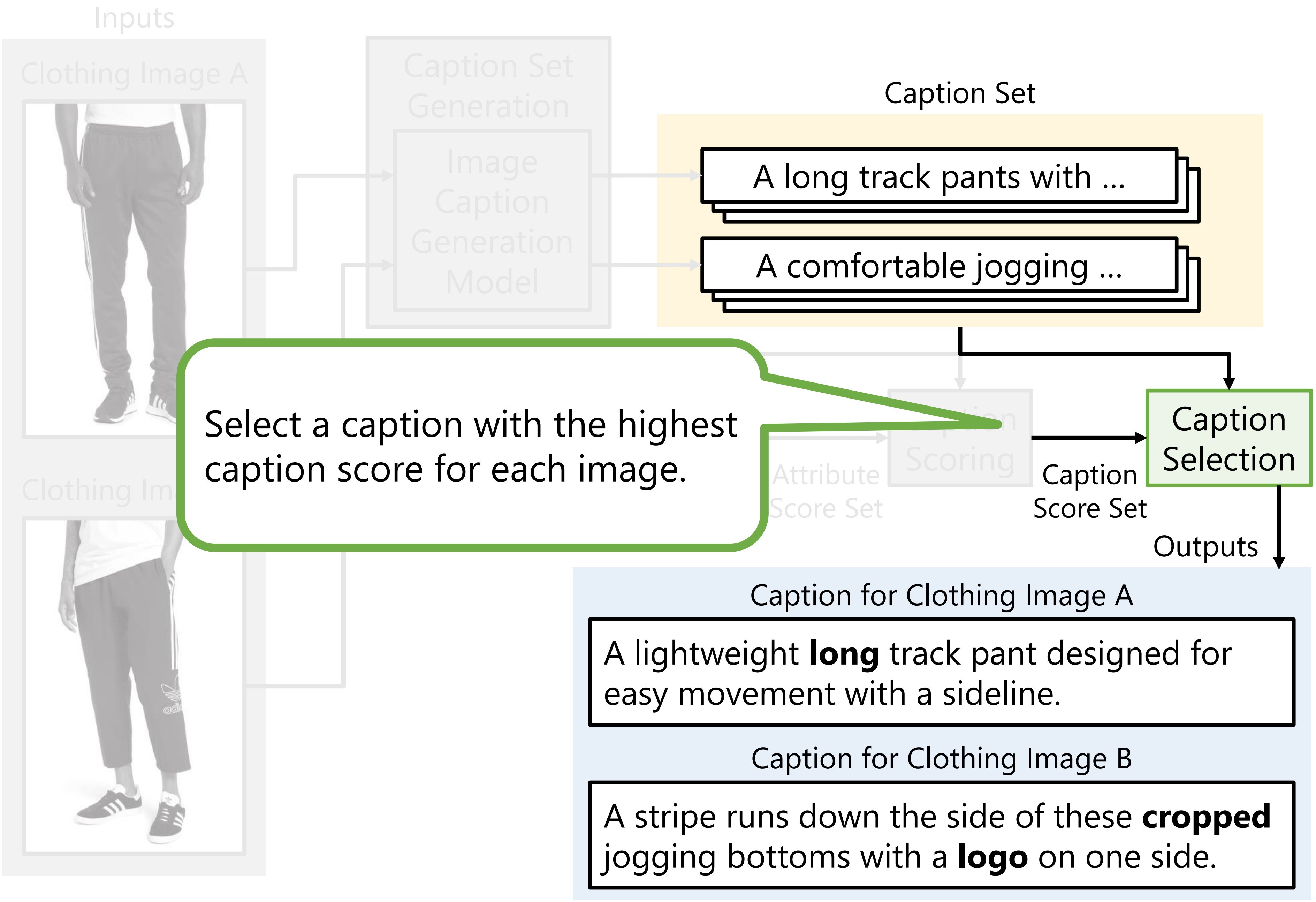
0.4

-0.5

②

Caption Score: $0.4 + (-0.5) = -0.1$

Caption Selection Module



Select a caption with the highest caption score for each image.

Caption for Clothing Image A

A lightweight **long** track pant designed for easy movement with a sideline.

Caption for Clothing Image B

A stripe runs down the side of these **cropped** jogging bottoms with a **logo** on one side.

Clothing Image Dataset

Comparison of Clothing Image Datasets

Datasets	Number of Images	Attributes	Captions
FACAD170K[10]	**178,849**	**yes**	**yes**
DeepFashion[11]	289,222	yes	no
FashionGen[12]	325,536	no	yes
iFashion[13]	1,062,550	yes	no

- A clothing image dataset with attributes and captions is required.
  - To train the image caption generation model
  - To evaluate the proposed method

- Adopted **FACAD170K**
  - Clothing images collected from websites
  - One-sentence caption for each image
  - 990 different attributes

A neat **band collar** is beautifully balanced by dropped **shoulder sleeve** in a staple **button** up top **cut** from **stretch** kissed **organic cotton**.

An example of data from FACAD170K (bold text indicates attributes)

[10] C. Cai, K.-H. Yap, and S. Wang, "Attribute conditioned fashion image captioning," IEEE International Conference on Image Processing, pp. 1921-1925, 2022.
[11] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.
[12] N. Rostamzadeh et al., "Fashion-gen: The generative fashion dataset and challenge," arXiv preprint arXiv:1806.08317, 2018.
[13] S. Guo et al., "The imaterialist fashion attribute dataset," Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0-0, 2019.

Preliminary Experiment

- Comparison and validation of the proposed method's multiple settings
  - Automatic evaluation of differently generated captions
- Validated settings
  - Attribute scoring module
    - **Using attribute frequency in caption sets as scores.**
    - Using estimated probabilities from a clothing attribute model as scores.
  - Caption scoring module
    - **Calculating and adding relative attribute scores for caption score.**
    - Comparing top n attribute scores between images, using differences in unique and common attribute occurrences for caption score.
- Evaluation methods
  - Precision, Recall, and F-measure between unique attributes annotated on one of the clothing images and attributes mentioned in captions
  - Generation of captions for 10,000 pairs and averaging each metric
- Results
  - Best performance with frequency-based attribute scoring and relative attribute score for caption scoring among the settings validated in this preliminary experiment.

Experiment

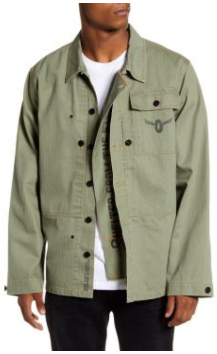
- Objectives
  - Evaluation of the captions generated by the proposed method in terms of:
    - How accurately they describe the features of a single clothing.
    - How well they describe the differences between clothing image pairs.
    - How useful the information provided is for comparing clothing.
  - Comparison of the effectiveness of the proposed method based on the similarity between clothing
- Methods
  - Presentation of 5 clothing image pairs and captions to subjects
    - Group of subjects: 10 men and women in their 20s.
  - 4 questions

Presented Clothing Image Pairs and Captions

Similar Pairs

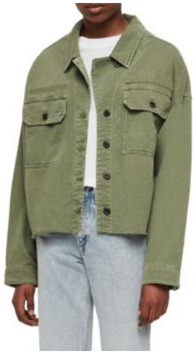
Pair 1

1A



Letter graphics add a military touch to a lightweight cotton-twill jacket.

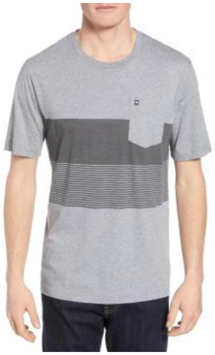
1B



Soft corduroy and a retro cropped hem give this slouchy, military-inspired denim jacket a lived-in edge.

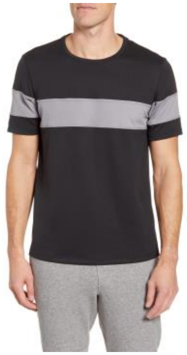
Pair 2

2A



A comfortable cotton T-shirt with chest pocket and horizontal stripes.

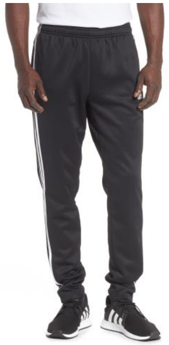
2B



Monotone chic tee with bold lines across the chest and sleeves.

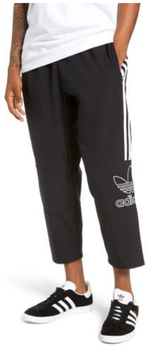
Pair 3

3A



A lightweight long track pant designed for easy movement with a sideline.

3B

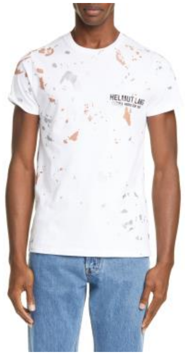


A stripe runs down the side of these cropped jogging bottoms with a logo on one side.

Dissimilar Pairs

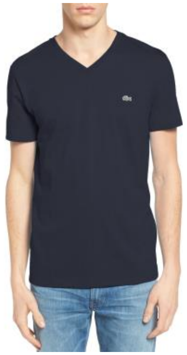
Pair 4

4A



Paint graphic to the front and letter logo to the chest of a comfort fit T-shirt.

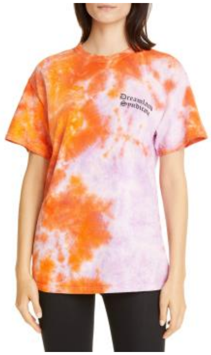
4B



A classic V-neck T-shirt in soft cotton with a signature logo at the chest.

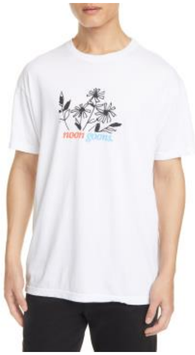
Pair 5

5A



A groovy tie-dye print washes over a crewneck tee that feels extra soft against your skin.

5B



The flower logo pops in vibrant color and dimension across the front of a cotton T-shirt.

- **Three similar pairs (Pair 1, 2, 3)** and **two dissimilar pairs (Pair 4, 5)**
  - Selected based on attribute matching and visual inspection

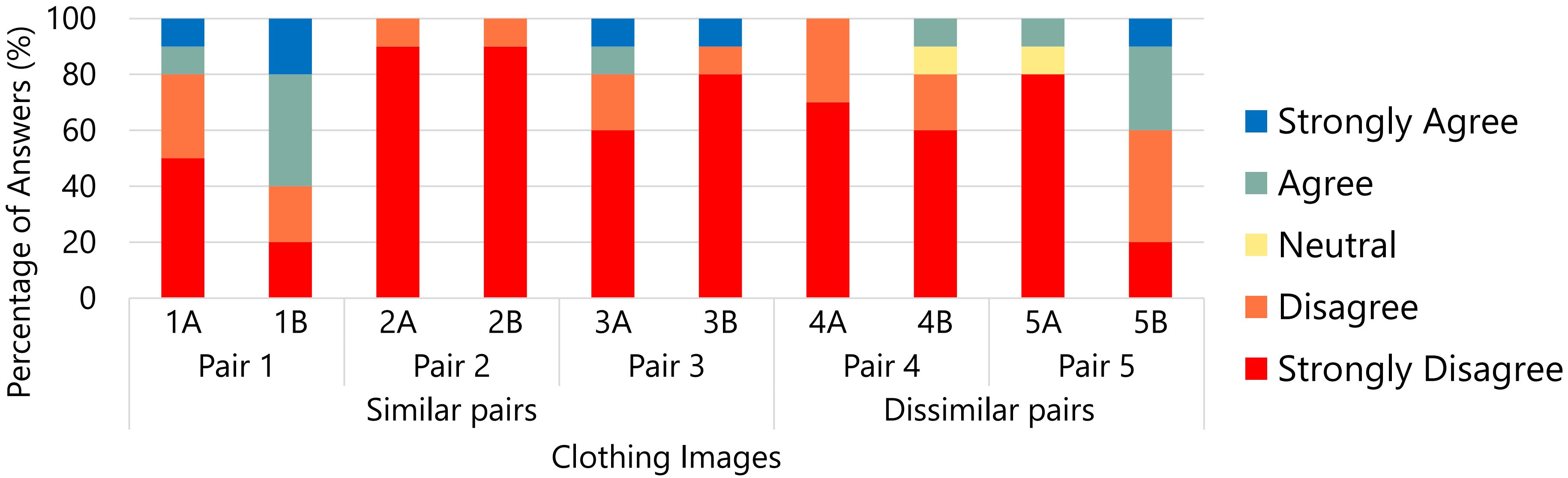
Questions and Test Method

- Set Questions and Options

	Questions	Options
Q1	Do you think the caption completely misdescribes a feature of the clothing?	• Strongly Agree • Agree • Neutral • Disagree • Strongly Disagree
Q2	Do you think the caption describes **one or more** features that are unique to the clothing?	
Q3	Do you think the caption describes **all** the features that are unique to the clothing?	
Q4	Do you think that the two captions would help you to compare the clothing if you were deciding whether to buy one of them?	

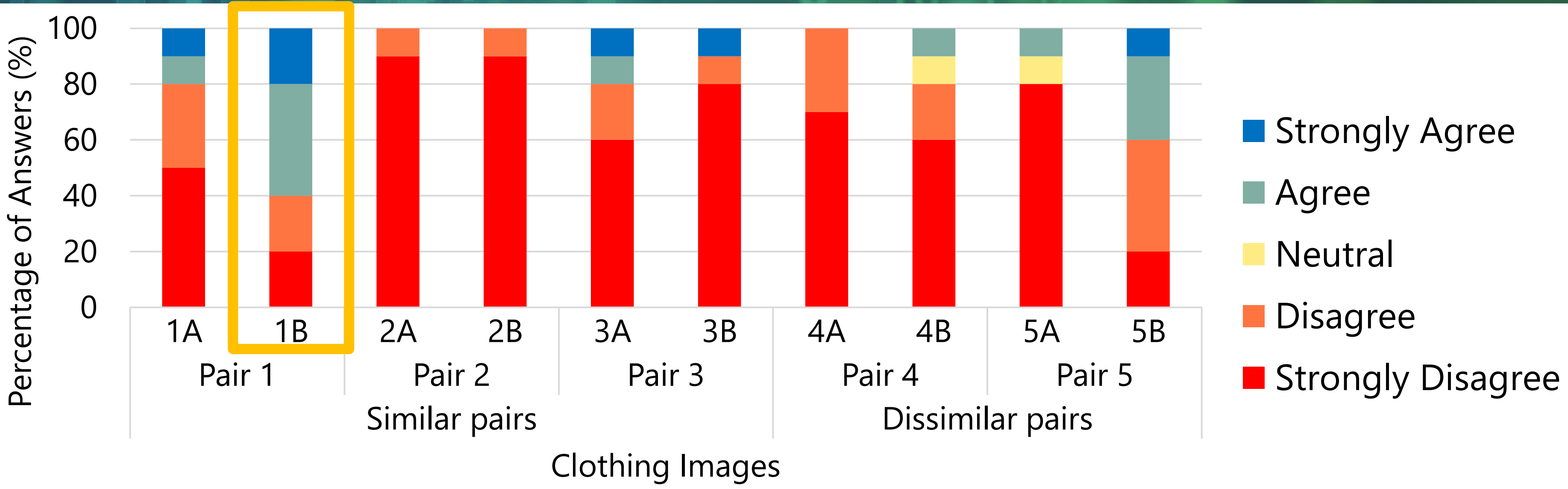
- Test Method
  - Wilcoxon Signed-Rank Test
    - Assessment of the significance of differences between answers in Q2 and Q3
  - Mann-Whitney U Test
    - Assessment of the significance of differences between answers to Q4 in similar and dissimilar pairs
  - Significance level: 5%
  - Bonferroni correction

Q1: Do you think the caption completely misdescribes a feature of the clothing?

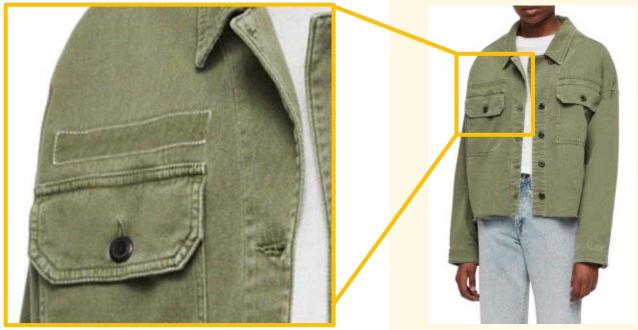


- Subjects answered "**No clear misdescriptions**" in many clothing images
  - This indicates that the captions generally accurately describe the features of the clothing.

Q1: Do you think the caption completely misdescribes a feature of the clothing?



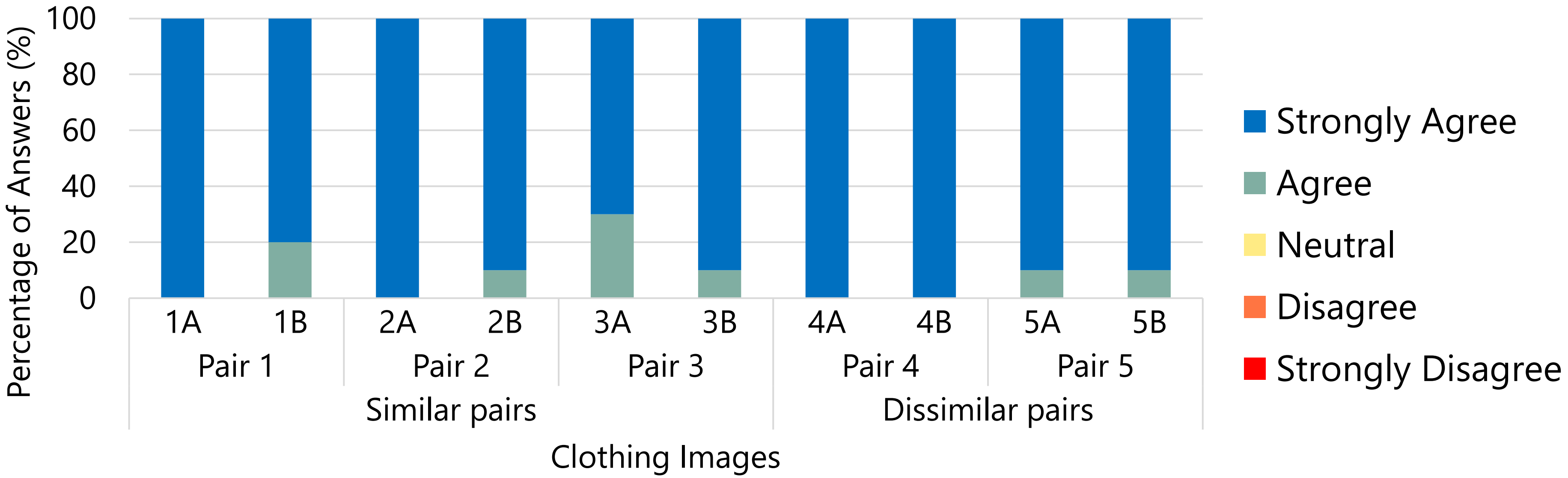
- Subjects answered "**No clear misdescriptions**" in many clothing images
  - This indicates that the captions generally accurately describe the features of the clothing.
- An example of misdescriptions
  - Denim material incorrectly described as "corduroy"
    - Likely due to difficulties in recognizing fine material details with CLIP.



Clothing image 1B

Soft **corduroy** and a retro cropped hem give this slouchy, military-inspired denim jacket a lived-in edge.

Q2: Do you think the caption describes one or more features that are unique to the clothing?

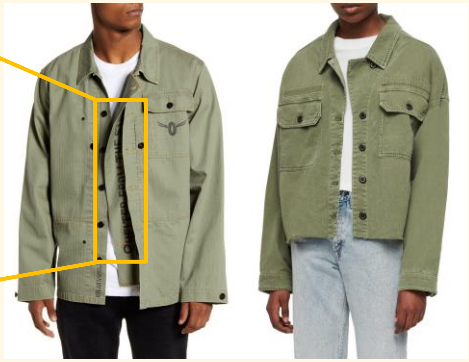


- All clothing images received **positive** answers.
  - This indicates that the captions describe at least one feature that is unique to the clothing.

Captions describing differences in graphics and length.

Clothing image 1A

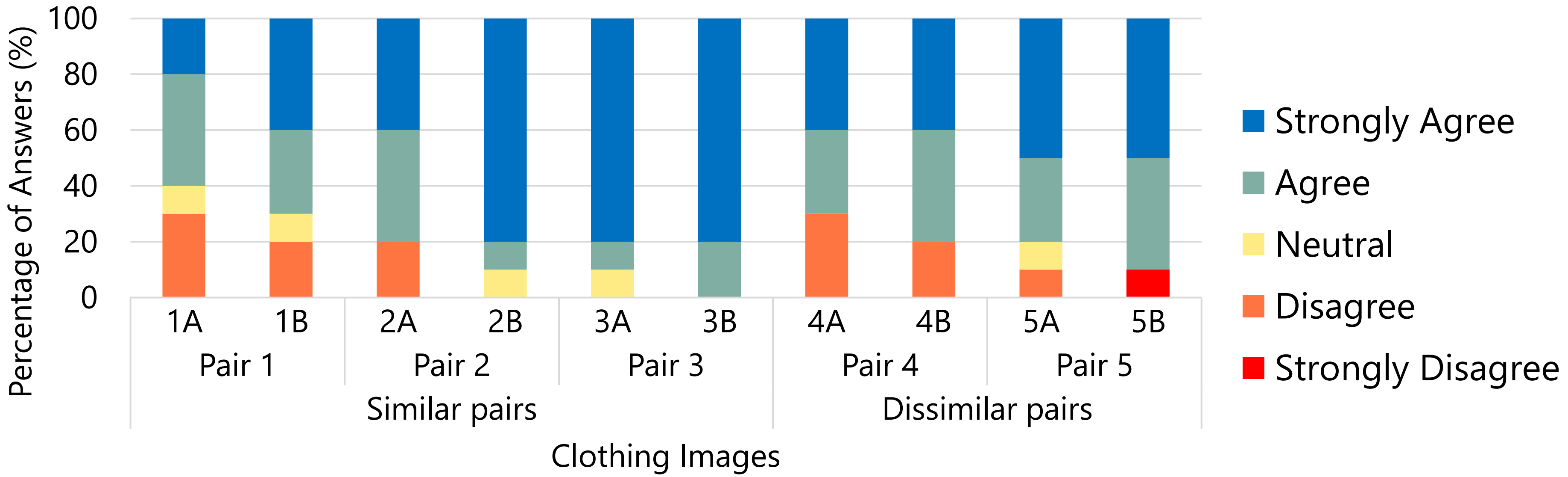
Letter graphics add a military touch to a lightweight cotton-twill jacket.



Clothing image 1B

Soft corduroy and a retro **cropped hem** give this slouchy, military-inspired denim jacket a lived-in edge.

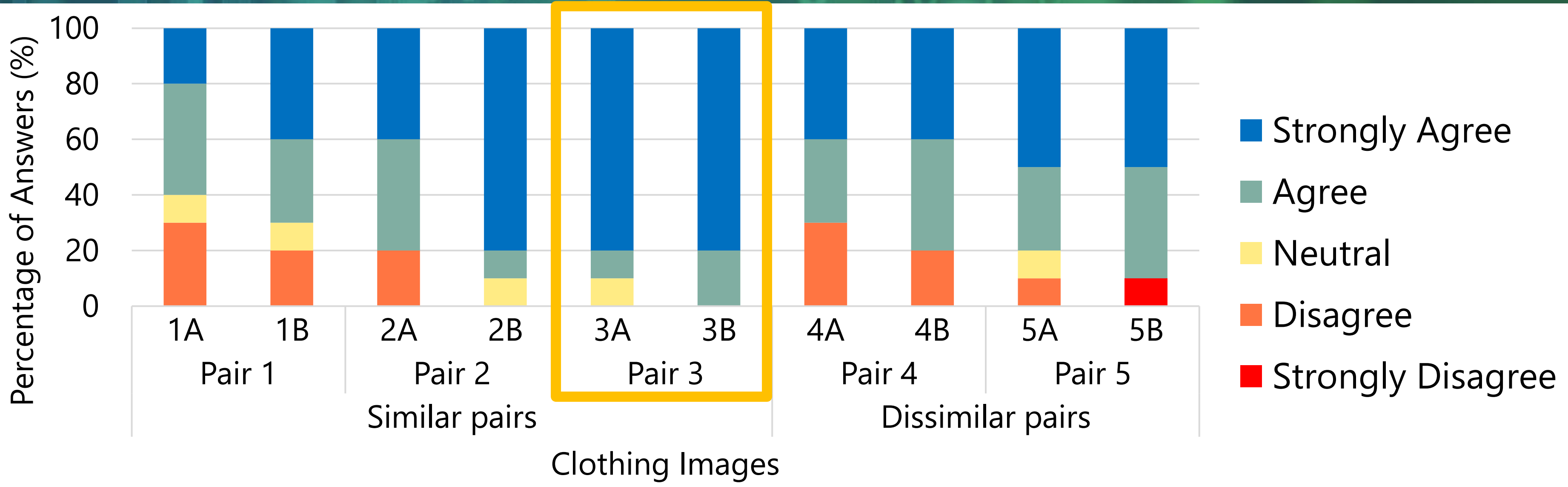
Q3: Do you think the caption describes all the features that are unique to the clothing?



- While many clothing received **positive** answers, the number of **negative** answers is higher than in Q2.
  - This indicates the captions often describe at least one unique feature, but they may not cover all such features.

p-value: $1.27 \times 10^{-8} < 0.025$

Q3: Do you think the caption describes all the features that are unique to the clothing?



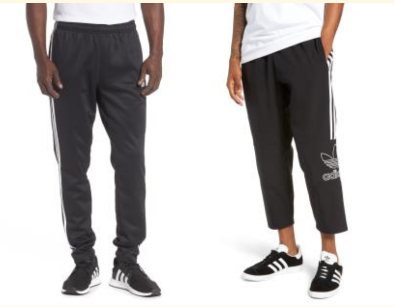
- While many clothing received **positive** answers, the number of **negative** answers is higher than in Q2.
  - This indicates the captions often describe at least one unique feature, but they may not cover all such features.

p-value: $1.27 \times 10^{-8} < 0.025$

- An example of pairs with many **positive** answers
  - Subjects comment: "Captions describe the length and the presence of the logo."

Clothing image 3A

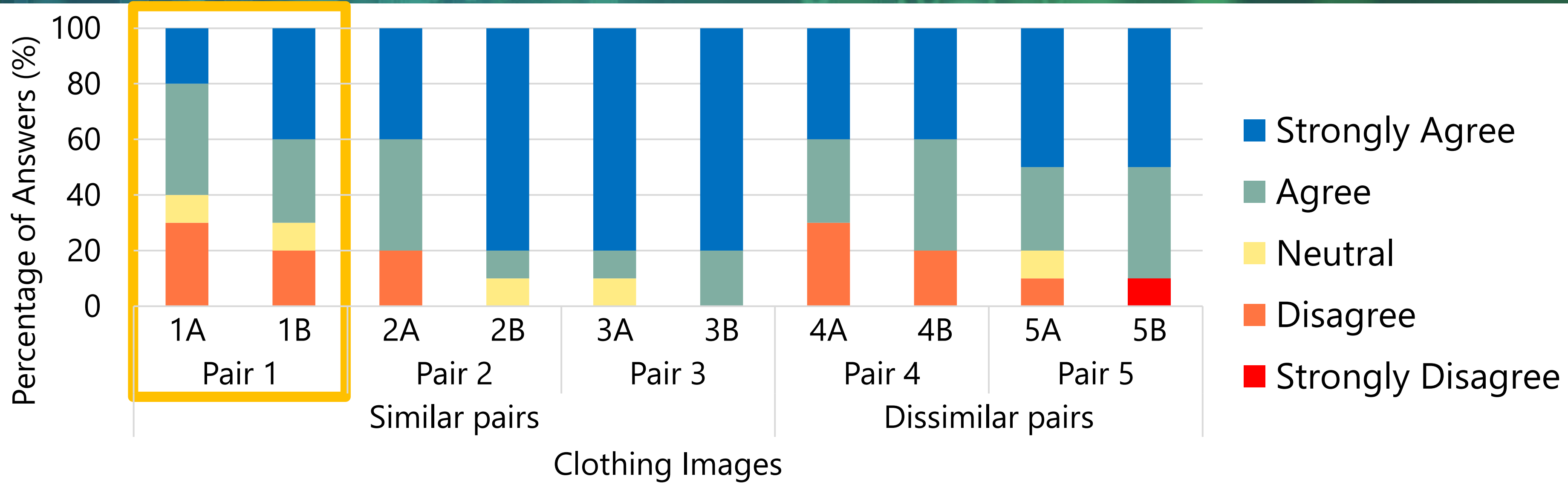
A lightweight **long** track pant designed for easy movement with a sideline.



Clothing image 3B

A stripe runs down the side of these **cropped** jogging bottoms with a **logo** on one side.

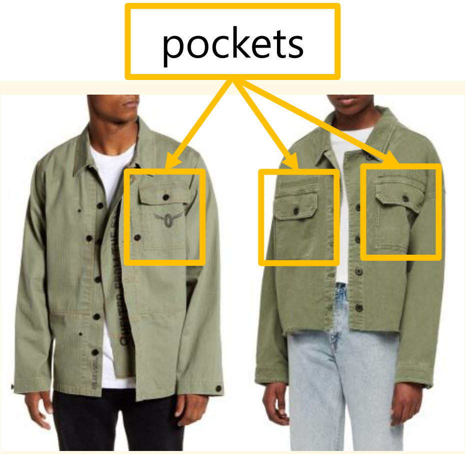
Q3: Do you think the caption describes all the features that are unique to the clothing?



- An example of pairs with many **negative** answers
  - Subjects comment: "The difference in the number of pockets wasn't described.“
  - While the attribute "pocket" can reflect the presence of pockets, it's difficult to describe the difference in their number.

Clothing image 1A

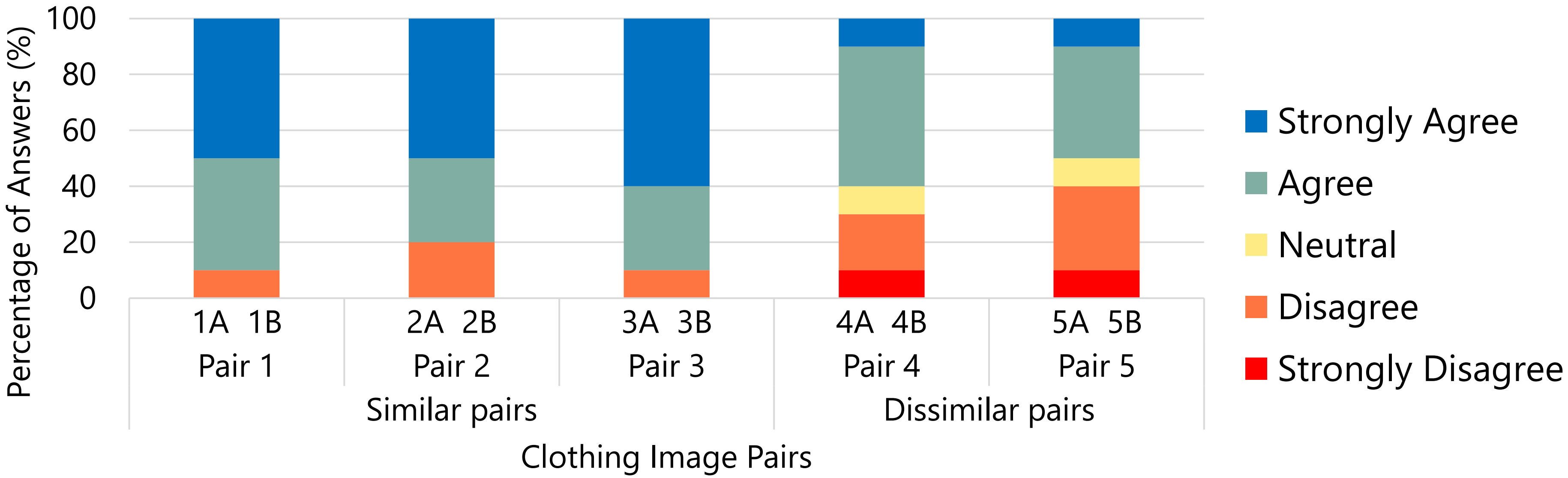
Letter graphics add a military touch to a lightweight cotton-twill jacket.



Clothing image 1B

Soft corduroy and a retro cropped hem give this slouchy, military-inspired denim jacket a lived-in edge.

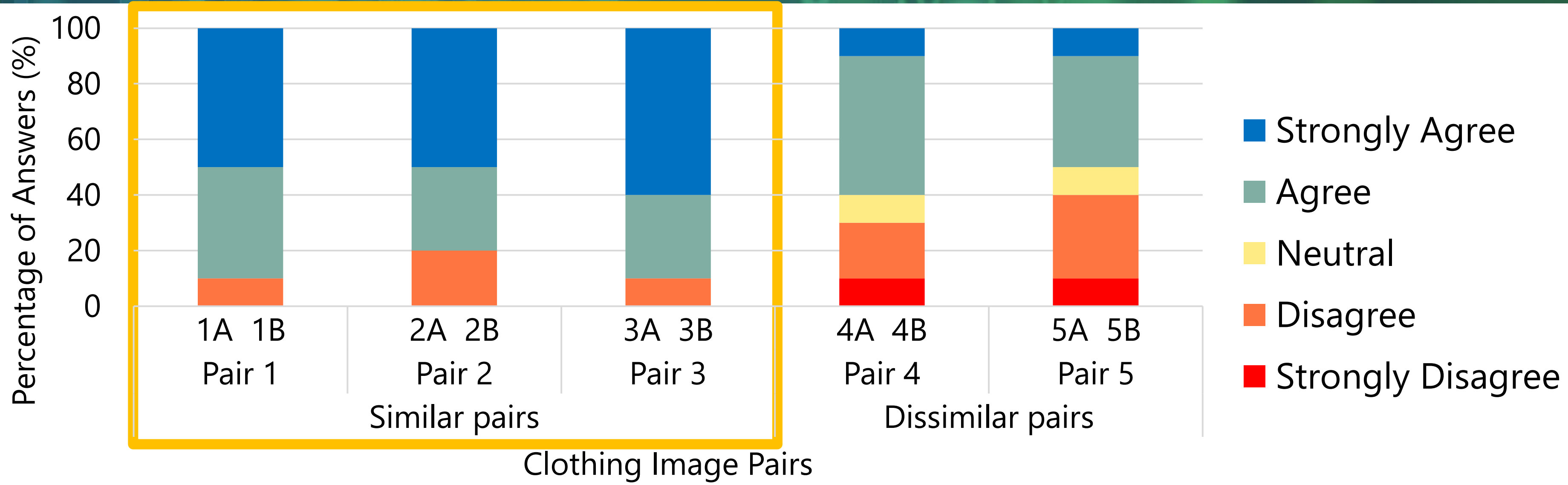
Q4: Do you think that the two captions would help you to compare the clothing if you were deciding whether to buy one of them?



- Similar pairs received more **positive** answers than dissimilar pairs.
  - This indicates the captions provide useful information for comparing pairs of highly similar clothing images.

p-value: $1.03 \times 10^{-3} < 0.025$

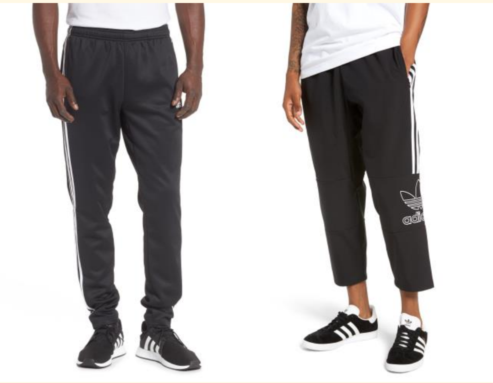
Q4: Do you think that the two captions would help you to compare the clothing if you were deciding whether to buy one of them?



- An example of **similar** pairs
  - Subjects comment: "By reading the captions, I was able to notice differences in length and logos.“
    - This indicates that the captions have the effect of drawing attention to differences in features that are difficult to notice and directly visible.

Clothing image 3A

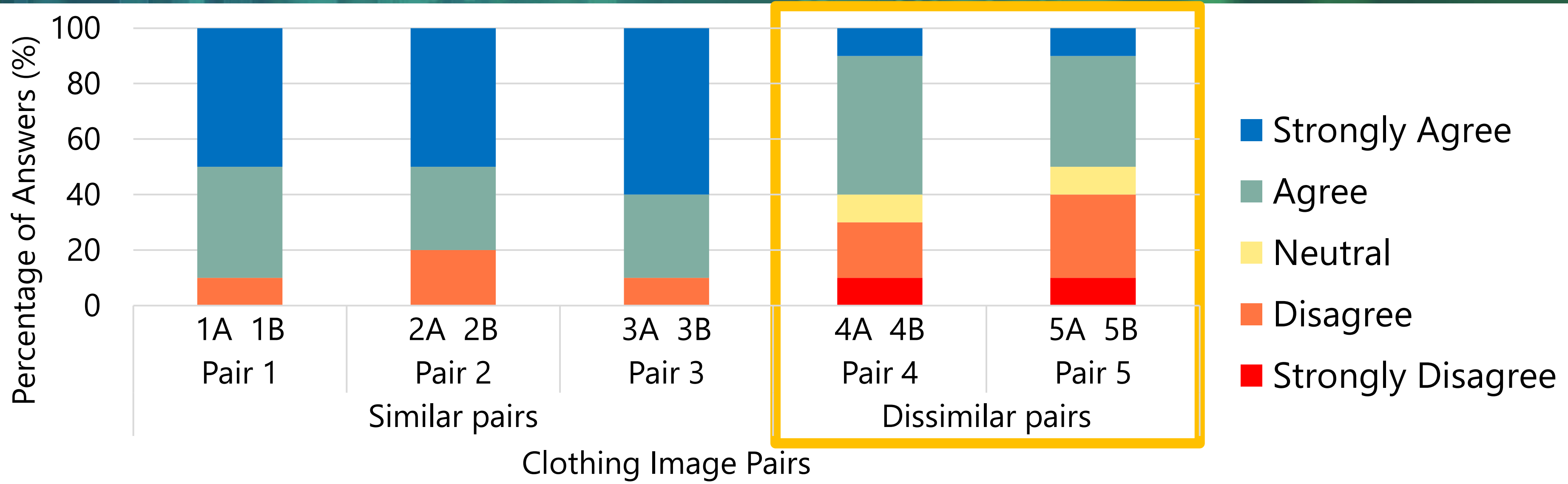
A lightweight **long** track pant designed for easy movement with a sideline.



Clothing image 3B

A stripe runs down the side of these **cropped** jogging bottoms with a **logo** on one side.

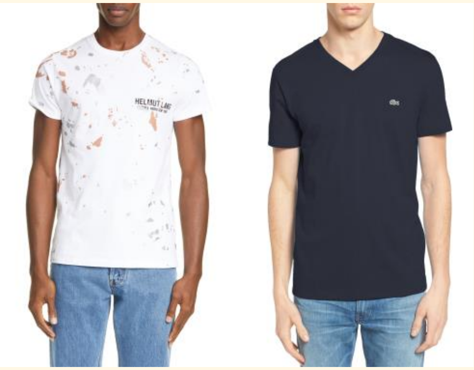
Q4: Do you think that the two captions would help you to compare the clothing if you were deciding whether to buy one of them?



- An example of **dissimilar** pairs
  - Subjects comment: "The differences were so apparent that the captions were unnecessary."
    - This indicates that the captions are less useful for comparison when the differences between the clothes are visually obvious.

Clothing image 4A

Paint graphic to the front and letter logo to the chest of a comfort fit T-shirt.



Clothing image 4B

A classic V-neck T-shirt in soft cotton with a signature logo at the chest.

Summary of Experiment

- Captions generated by the proposed method
  - Generally, accurately describe the features of clothing images.
  - Often describe at least one unique feature but may not always cover all such features.
  - Provide useful information for comparing clothing pairs with high similarity.
    - Especially for subtle differences that are not visually obvious.
  - Effectiveness is lower for pairs with low similarity, where differences are visually clear.

Conclusion

- Development of a system to provide relevant information to consumers comparing clothing
  - Generation of captions highlighting the differences between a clothing image pair
    - Generation of multiple captions for each clothing image
    - Calculation of **attribute scores** based on the frequency of attributes in the caption set
    - Calculation of **caption scores** by sum the relative attribute scores of the attributes included in the captions
    - Selection of the caption with the highest caption score
- Results of the experiment
  - The proposed method can provide useful information for comparing high similarity clothing image pairs.
- Future work
  - Extension of the approach to enable comparisons among three or more clothing images