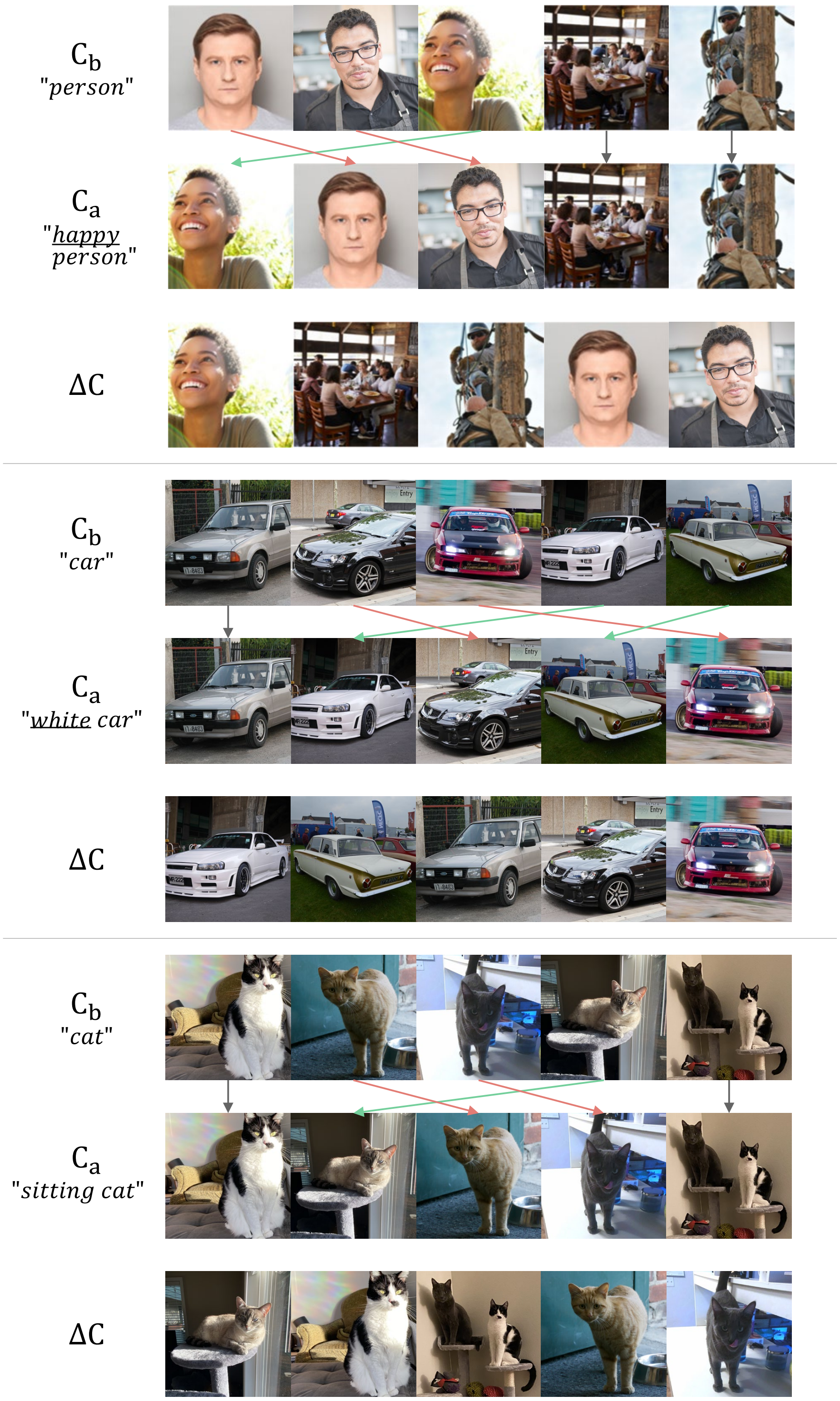


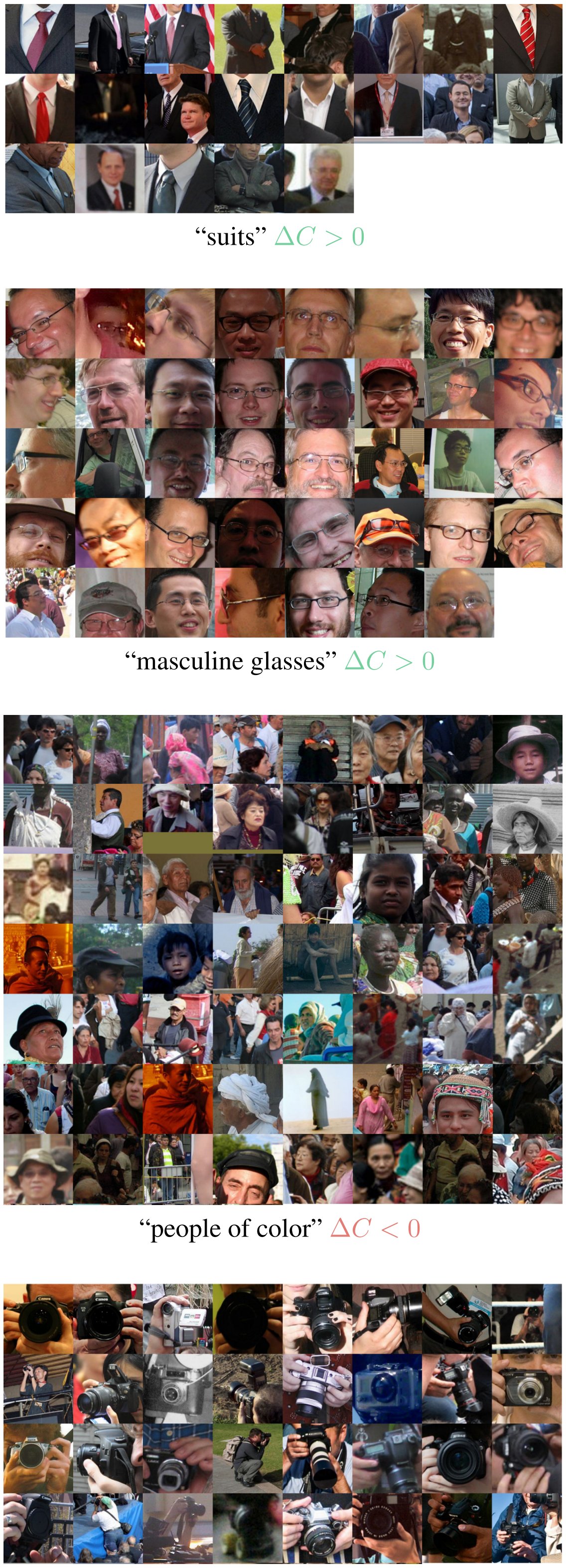
Figure 1: Additional ΔC ranking examples where C_b and C_a are the baseline and augmented captions, respectively.

Question	Response Type	
Do you consent to the study?	yes / no	
What is your age?	free response	
What is your gender?	select all that apply	(e.g. *non-binary, woman)*
What are your pronouns?	select all that apply	(e.g. *they/them, he/him)*
What is your most recent degree program?	free response	
Do you have at least two years of professional AI/ML experience?	free response	
Have you taken three or more AI/ML courses?	yes / no	
Please list all AI/ML related courses.	free response	
What is your expertise level in AI/ML?	scale (0 - 5)	
Do you have ViL experience?	select all that apply	(e.g. *ViL navigation, VQA)*
Describe your experience with the one(s) above.	free response	
Have you used any tools or libraries for analyzing ViL behavior?	yes / no	
Which of the following tools/libraries have you used?	select all that apply	(e.g. *TensorBoard, matplotlib)*
Can you tell us why you used it and for what purpose?	free response	

Table 1: Pre-study questions and response types given before the interface tutorial.

Question	Response Type
The tool was easy to learn how to use.	Likert (1 - 7)
The tool was easy to use.	Likert (1 - 7)
I felt confident when using the tool.	Likert (1 - 7)
I enjoyed using the tool.	Likert (1 - 7)
I would like to use a tool like this one again.	Likert (1 - 7)
I am confident the image sets I created with this tool capture my intent.	Likert (1 - 7)
This tool is helpful for finding new model behavior.	Likert (1 - 7)
This tool is helpful for confirming my understanding of model behavior.	Likert (1 - 7)
It was easy to build sets of images capturing a concept I was looking for.	Likert (1 - 7)
It was easy to find additional relevant images to add to my image sets.	Likert (1 - 7)
The images within sets I created are coherent with each other.	Likert (1 - 7)
The image sets I created capture a systemic biased relationship between inputs to the model.	Likert (1 - 7)
What was your favorite part of using the tool?	free response
What was the most frustrating part of using the tool?	free response
Are there any other comments you have about this tool?	free response

Table 2: Post-study questions and response types given after the participant has completed both tasks.



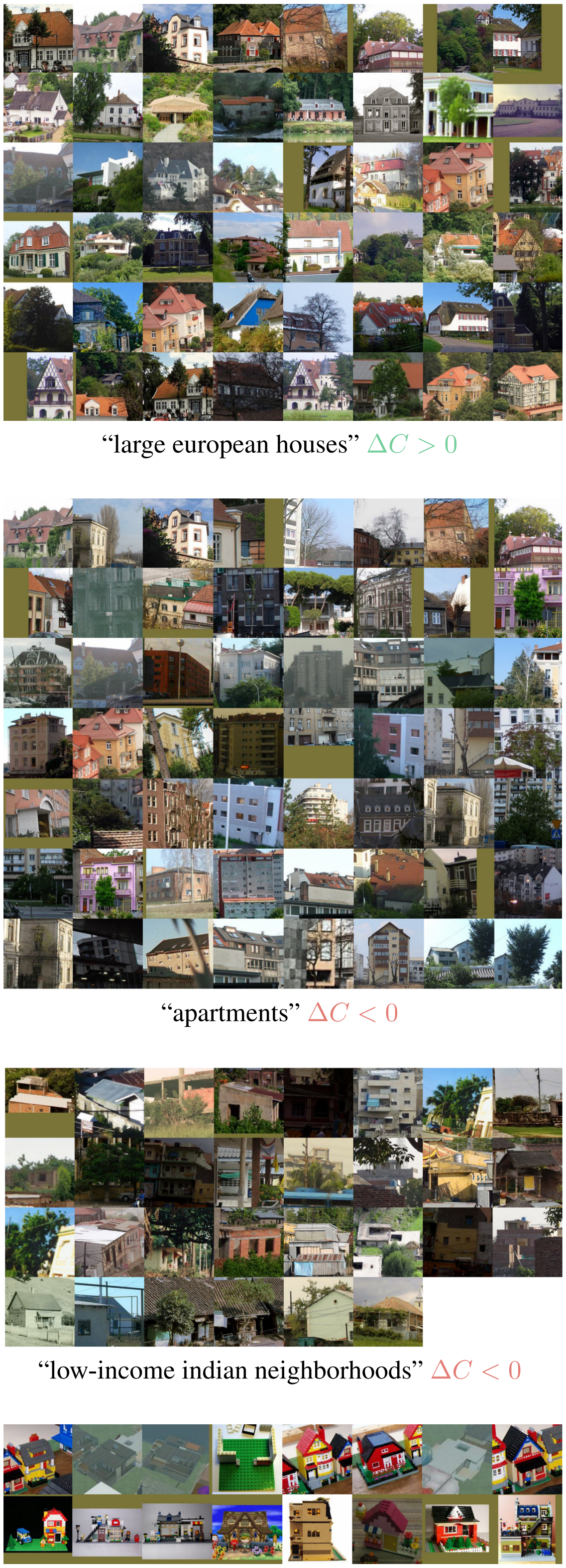
“suits” $\Delta C > 0$

“masculine glasses” $\Delta C > 0$

“people of color” $\Delta C < 0$

“cameras” $\Delta C \approx 0$

Figure 2: Additional example slices created by participants for the Person/CEO task with VL*Slice*.



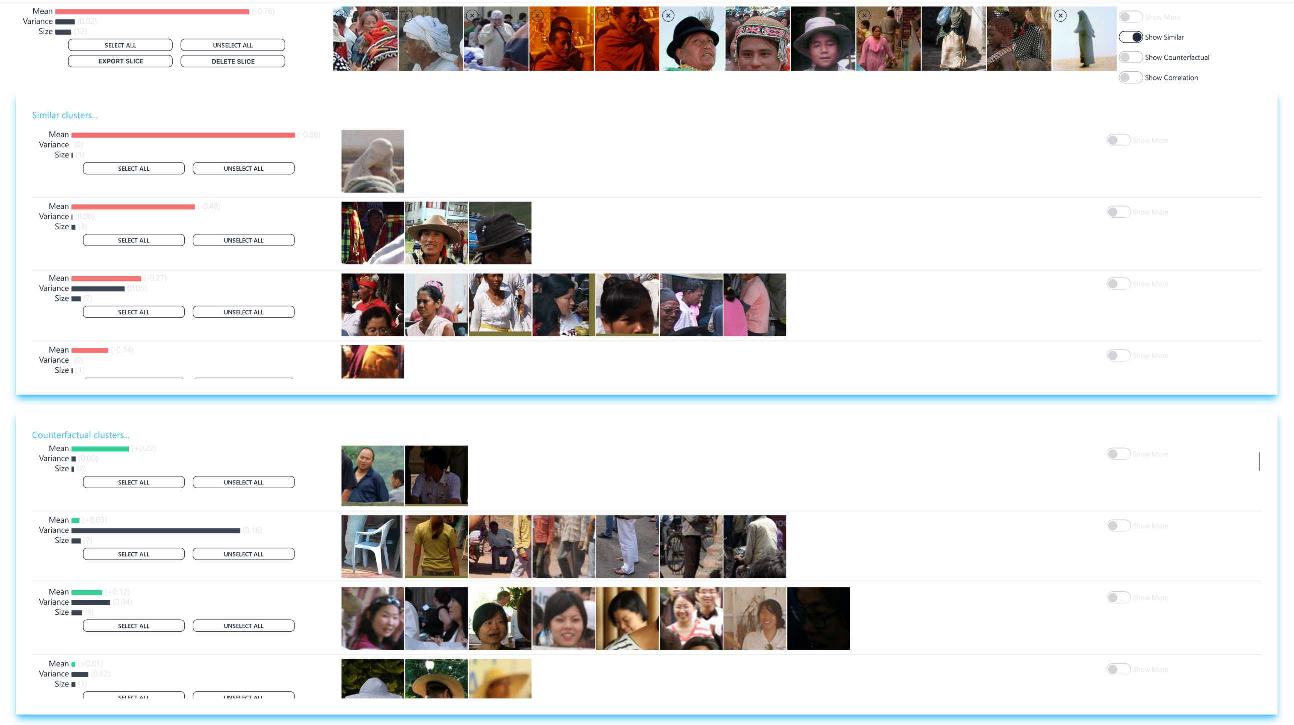
“large european houses” $\Delta C > 0$

“apartments” $\Delta C < 0$

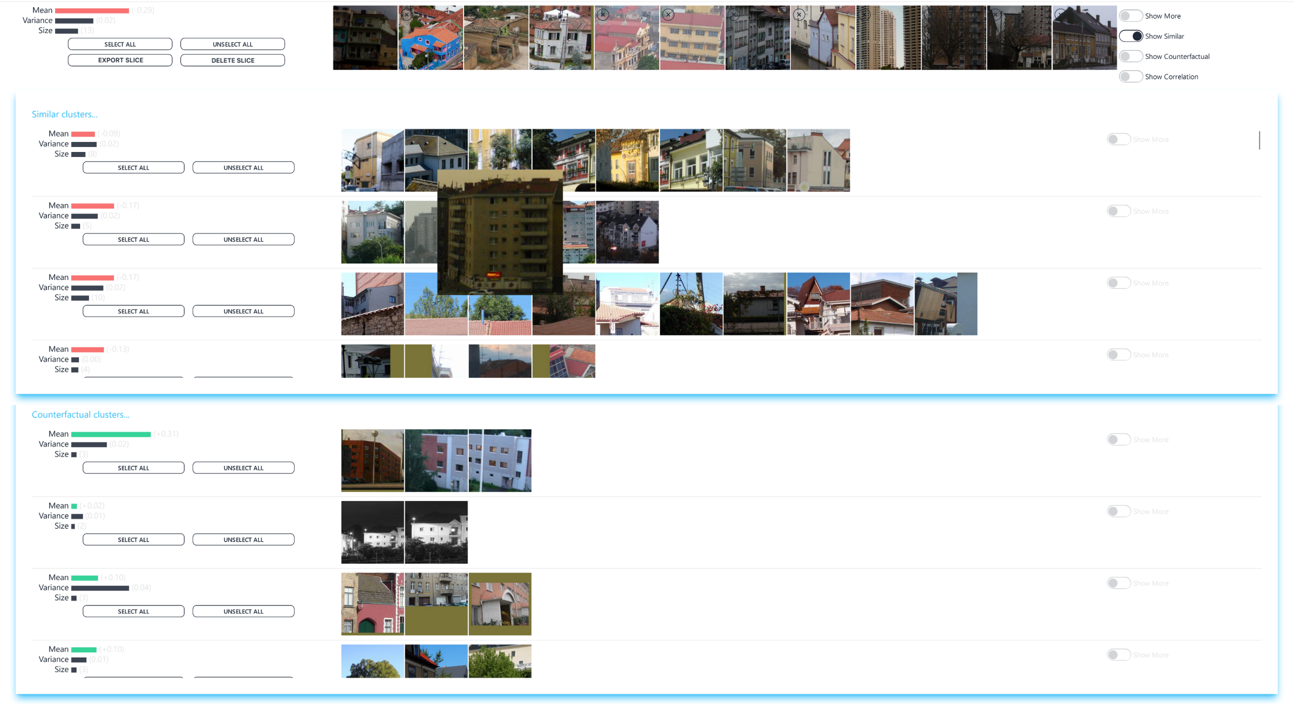
“low-income indian neighborhoods” $\Delta C < 0$

“legos” $\Delta C \approx 0$

Figure 3: Additional example slices created by participants for the House/Nice House task with VL*Slice*.

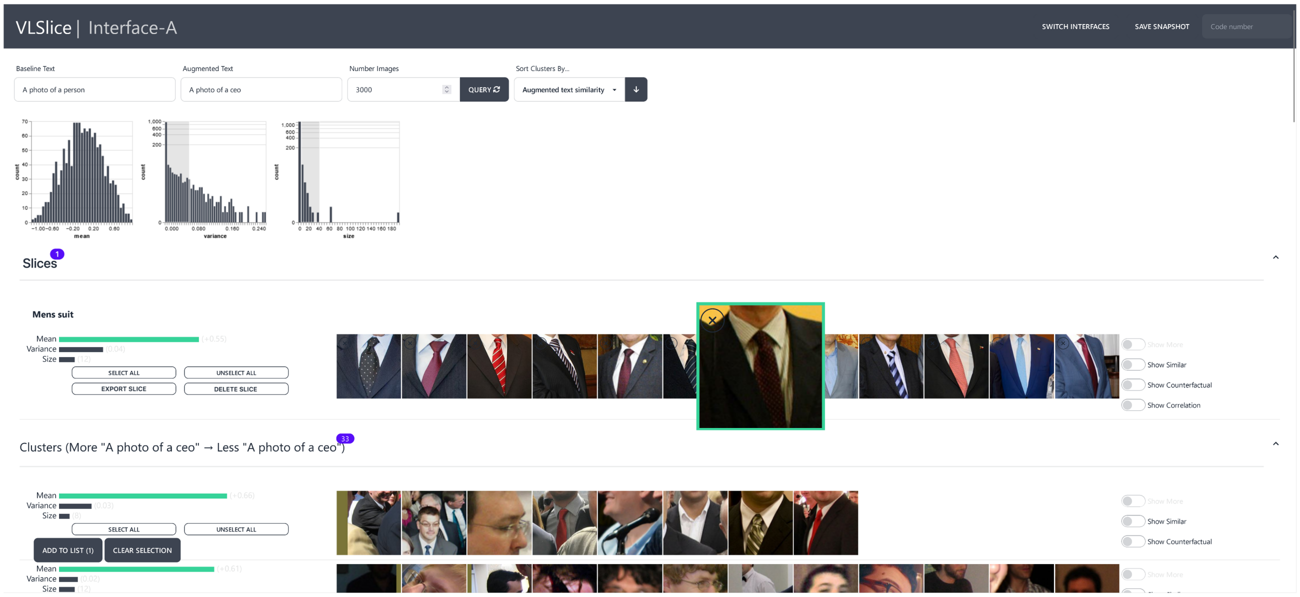


“non-western clothing” $\Delta C < 0$



“apartments” $\Delta C < 0$

Figure 4: `VL`*Slice* similar and counterfactual cluster recommendation interface examples.



VLSlice | Interface-A

Baseline Text: A photo of a person

Augmented Text: A photo of a ceo

Number Images: 3000

Sort Clusters By...: Augmented text similarity

Slices 1

Mens suit

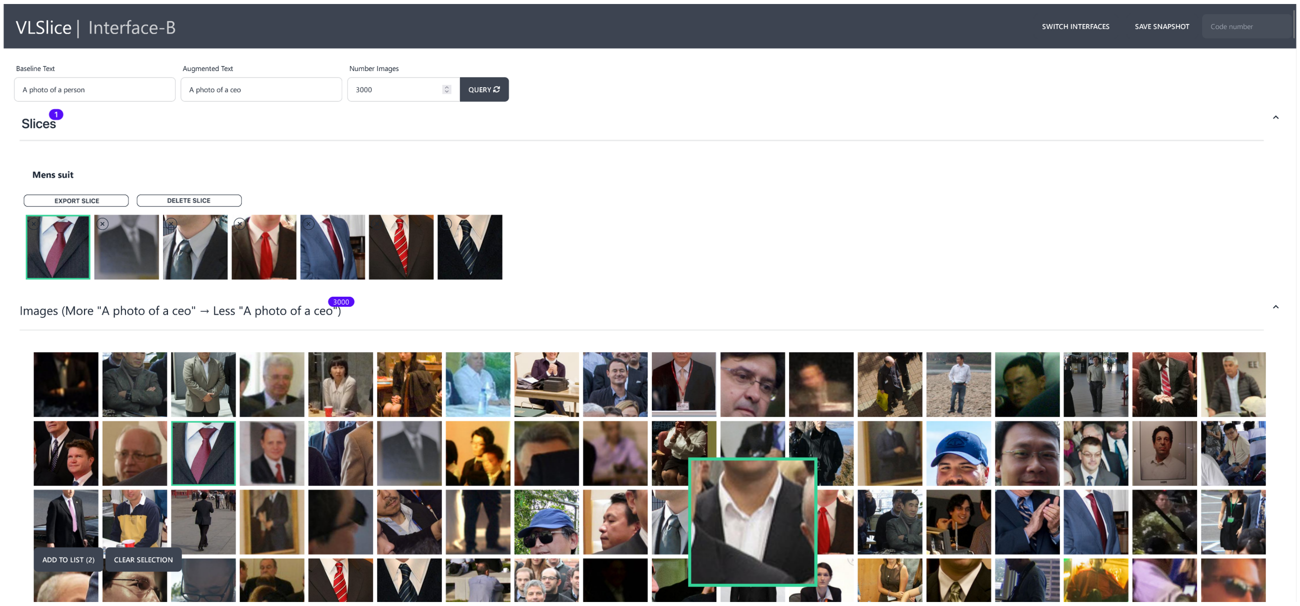
Mean (+0.55)
Variance (0.04)
Size (12)

Clusters (More "A photo of a ceo" → Less "A photo of a ceo") 33

Mean (+0.66)
Variance (0.03)
Size (8)

Mean (+0.61)
Variance (0.02)
Size (12)

Figure 5: `VL`*Slice* interface screenshot. Clicking "*show similar*" or "*show counterfactual*" expands to display recommendations like those shown in Fig. 4



VLSlice | Interface-B

Baseline Text: A photo of a person

Augmented Text: A photo of a ceo

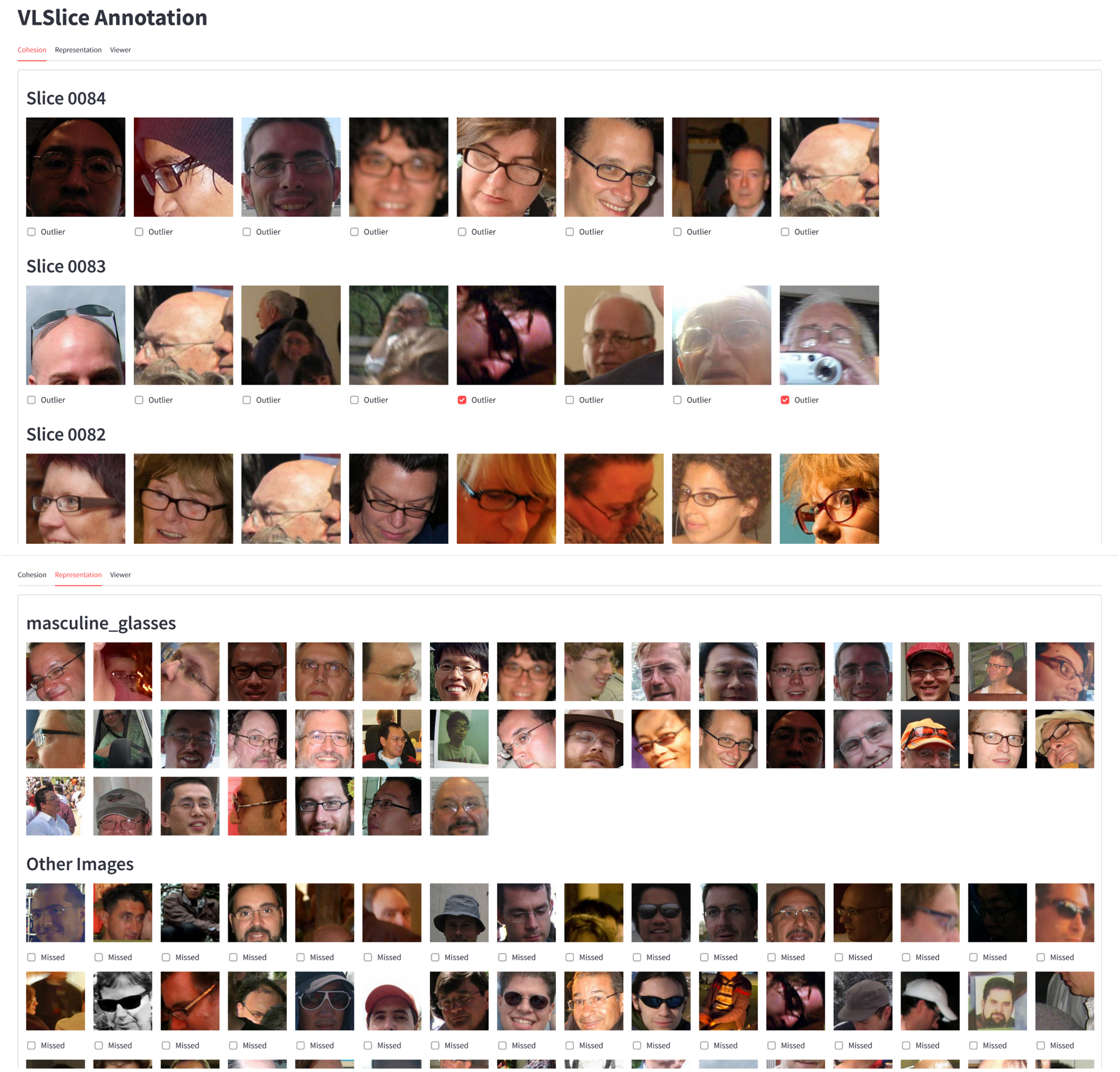
Number Images: 3000

Slices 1

Mens suit

Images (More "A photo of a ceo" → Less "A photo of a ceo") 3000

Figure 6: ListSort interface screenshot.



VLSlice Annotation

Slice 0084

☐ Outlier ☐ Outlier ☐ Outlier ☐ Outlier ☐ Outlier ☐ Outlier ☐ Outlier ☐ Outlier

Slice 0083

☐ Outlier ☐ Outlier ☐ Outlier ☐ Outlier ☑ Outlier ☐ Outlier ☐ Outlier ☑ Outlier

Slice 0082

masculine_glasses

Other Images

☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed

☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed ☐ Missed

Figure 7: Annotation interface for cohesion (top) and representation (bottom). Annotators select all outlier images for a slice in the first case and any missed images for a slice in the second. No annotator sees the same slice across tasks.