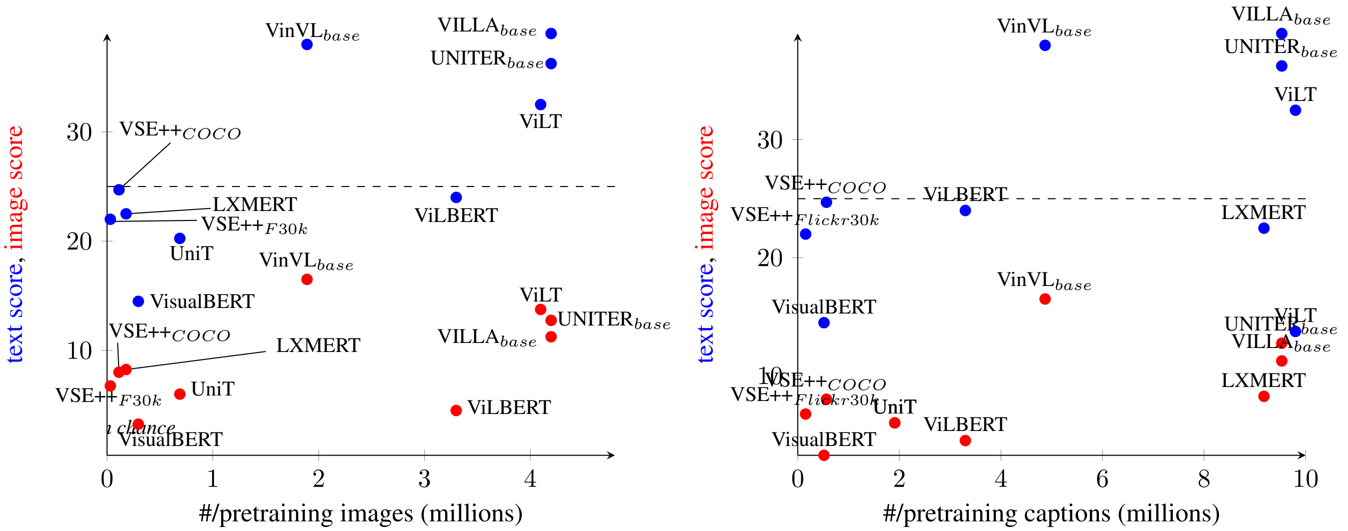
A. Confidence Intervals

Model	Text		Image		Group	
MTurk Human	89.50	[88.58,90.42]	88.50	[85.74,91.26]	85.50	[82.45,88.55]
VinVL	37.75	[29.81,45.69]	17.75	[12.03,23.47]	14.50	[10.71,18.29]
UNITER$_{large}$	38.00	[32.05,43.95]	14.00	[9.89,18.11]	10.50	[8.45,12.55]
UNITER$_{base}$	32.25	[24.10,40.40]	13.25	[7.53,18.97]	10.00	[7.09,12.91]
ViLLA$_{large}$	37.00	[31.34,42.66]	13.25	[5.63,20.87]	11.00	[5.97,16.03]
ViLLA$_{base}$	30.00	[21.99,38.01]	12.00	[8.56,15.44]	8.00	[4.56,11.44]
VisualBERT	15.50	[12.74,18.26]	2.50	[0.45,4.55]	1.50	[0.00,3.55]
ViLT	34.75	[27.47,42.03]	14.00	[11.09,16.91]	9.25	[6.53,11.97]
LXMERT	19.25	[13.83,24.67]	7.00	[3.56,10.44]	4.00	[0.56,7.44]
ViLBERT	23.75	[15.19,32.31]	7.25	[5.25,9.25]	4.75	[1.47,8.03]
UniT	19.50	[16.19,22.81]	6.25	[2.07,10.43]	4.00	[0.56,7.44]
CLIP	30.75	[21.90,39.60]	10.50	[5.91,15.09]	8.00	[4.56,11.44]
VSE++$_{COCO}$ (ResNet)	22.75	[19.47,26.03]	8.00	[5.09,10.91]	4.00	[2.70,5.30]
VSE++$_{COCO}$ (VGG)	18.75	[13.82,23.68]	5.50	[2.74,8.26]	3.50	[0.74,6.26]
VSE++$_{Flickr30k}$ (ResNet)	20.00	[15.32,24.68]	5.00	[0.00,10.51]	2.75	[0.03,5.47]
VSE++$_{Flickr30k}$ (VGG)	19.75	[14.49,25.01]	6.25	[1.16,11.34]	4.50	[1.45,7.55]
VSRN$_{COCO}$	17.50	[11.62,23.38]	7.00	[4.09,9.91]	3.75	[2.23,5.27]
VSRN$_{Flickr30k}$	20.00	[16.10,23.90]	5.00	[2.75,7.25]	3.50	[0.00,7.29]

Table 1. 95% confidence intervals for the aggregate results on Winoground. We divided the dataset into 4 groups of equal size to get 4 scores for each model and score-type, and used this to compute the confidence intervals.

B. Impact of Pretraining Data

In investigating the impact of pretraining data, we found that the number of pretraining images correlates with higher text scores, even though all models still perform poorly (see below). Interestingly, the image scores do not show the same correlation as the text scores, which we plan to explore more in future work. We observe the same general trend for the text and images score versus the number of pretraining captions.



C. Linguistic Tag Breakdown

Tag	Fine-Grained Tag	Example
	Noun Phrase, Determiner-Numeral	[a person] carrying [more than one flotation device]
	Noun Phrase	[a person] holding up [books]
	Determiner-Numeral, Noun Phrase	[a lightbulb] surrounding [some plants]
Object	Noun Phrase, Determiner-Possessive	[a deer's nose] is resting on [a child's hand]
	Noun Phrase, Adjective-Color	aerial view of a green tree in [the brown freshly turned soil] next to [a green field]
	Pronoun, Noun Phrase	[the person] wears a hat but [it] doesn't
	Determiner-Numeral Phrase	[one] is in a boat and [almost everyone] is swimming
	Pronoun, Verb-Intransitive	[it] ran away while [they] pursued
	Noun	more [bicycles] than [cars]
	Adjective-Age	[an older] person blocking [a younger] person
	Scope, Preposition	racing [over] it []
	Verb-Intransitive, Verb-Transitive Phrase	a kid [threw a basketball] then [jumped]
	Verb-Intransitive, Adjective-Manner	the younger person is [making noise] while the other is [silent]
	Negation, Noun Phrase, Preposition Phrase	a person [with long braids] is exercising in front of a person [without braids]
	Scope, Preposition, Verb-Intransitive	[out]1[swam]2 the person in the red swimcap []2[]1
	Noun Phrase, Adjective-Animate	the one on the left is [sad] and the other is [happy]
	Adjective-Size	the [taller] person hugs the [shorter] person
	Determiner-Possessive	the [person's] leg is on the [dog's] torso
	Adjective-Texture	[smooth] shoes are on a [soft] floor
	Adjective-Color	painting the [white] wall [red]
	Scope	[getting] a horse [] wet
	Preposition Phrase	flat [at the bottom] and pointy [on top]
	Relative Clause, Scope	the person [who is wearing a crown] is kissing a frog []
	Adjective-Height	a [taller] person wearing blue standing next to a [shorter] person
	Verb-Intransitive Phrase, Preposition	the gesture of the person [sitting down] is supporting the understanding of the person [standing up]
	Verb-Intransitive, Determiner-Numeral	some people are [standing] but more are [sitting]
	Determiner-Numeral	[one]1 person[]2 wearing [two]1 scarf[s]2
	Adjective-Weight	the larger ball is [lighter] and the smaller one is [heavier]
	Verb-Intransitive, Noun	the dog is [standing] and the person is [swimming]
	Verb-Intransitive Phrase, Adverb-Animate	the person on the left is [crying sadly] while the one on the right is [smiling happily]
	Scope, Relative Clause	a fencer [who is wearing black pants] having a point scored against them by another fencer [] using a wheelchair
	Adjective-Speed	the train is [still] while the person is [moving fast]
	Adverb-Temporal	a person is drinking [now] and eating [later]
	Adverb-Spatial	the car is sitting [upside down] while the person is standing [rightside up]
Relation	Adjective-Shape	the [round] table has a [square] base
	Noun, Adjective-Color	Young person playing baseball with a [blue] bat and [green] ball
	Verb-Transitive	the person with the ponytail [buys] stuff and other [packs] it
	Scope, Verb-Transitive	[] gears for [moving] something
	Scope, Preposition Phrase	[] child in [front facing] row of yellow rubber ducks
	Adjective-Temperature	a [hot] drink on a [cold] day
	Adjective-Temporal	the [first] vowel is E and the [last] consonant is N
	Scope, Conjunction	a person spraying water on [someone else]1 [and]2 a person on a bike []2 []1
	Scope, Conjunction Phrase	A child [] riding a bike [and an adult]
	Preposition Phrase, Scope	someone [with an apple] is hurt by a tree []
	Adjective-Manner Phrase	two people wearing clothes of [different] colors are on [the same] side of the tennis net
	Verb-Intransitive	a person [stands] and a dog [sits]
	Adjective-Animate	[toy] cat with [real] baby
	Adverb-Spatial Phrase	the sailboat sails [close] but the beach is [far away]
	Scope, Adjective-Texture	A [] small animal with [curled] hair
	Adverb-Animate	someone talks on the phone [angrily] while another person sits [happily]
	Adjective-Manner	[poor] [unfortunate] people
	Verb-Transitive Phrase	they [drank water] then they [worked out]
	Adjective-Color (3-way swap)	The [red]→[yellow] book is above the [yellow]→[blue] book and below the [blue]→[red] book
	Scope, Adjective-Manner	[] living things [drinking]
	Preposition	seat numbers increasing from [right] to [left]
	Verb-Intransitive Phrase	a cat is [stretching] and a person is [lying down]
	Sentence	[the coffee is poured] before [it is ground]
	Adjective-Speed Phrase, Verb-Intransitive	the person with green legs is running [quite slowly] and the red legged one runs [faster]
	Adjective-Spatial	A [left] hand pulls a glove onto a [right] hand
	Negation, Scope	The [un]caged bird has an []opened cage door
	Verb-Transitive Phrase, Verb-Intransitive, Preposition Phrase	the dog [bite]1s []2 what someone would normally [wear]1 [as a hat]2
	Altered POS	[watch]ing the [present]
	Verb-Transitive, Noun	someone []1 on [the ground]2 [is]1 spraying water towards [a vehicle]2
	Scope, Altered POS, Verb-Intransitive, Verb-Transitive	[walking]1 someone []1 [cut]2 [lines]2 into green plants
	Noun, Adjective-Size	the [person]1 is too [big]2 for the [small]2 [door]1
Both	Noun, Verb-Intransitive	a [dog sitting] on a couch with a [person lying] on the floor
	Scope, Noun, Preposition	[]1 a person [near]1 [water]2 using a []2 lasso
	Noun, Preposition Phrase, Scope	a person wearing a [bear]1 mask []2 in blue on the left hand side of a person wearing a [panda]1 mask [with glasses]2 in pink
	Scope, Preposition Phrase, Adjective-Color	[darker]1 things []2 become [light]1 [in stripes]2
	Altered POS, Determiner-Numeral	[one] ear that some [donkey] is whispering a secret into

Table 2. Examples showcasing the full linguistic (swap-dependent) tag breakdown.

D. Heatmaps for the Word-Region Alignment Models

We provide heatmaps for models that use word-region alignment: UNITER, ViLLA and ViLT. See the main text for the ViLT heatmaps.

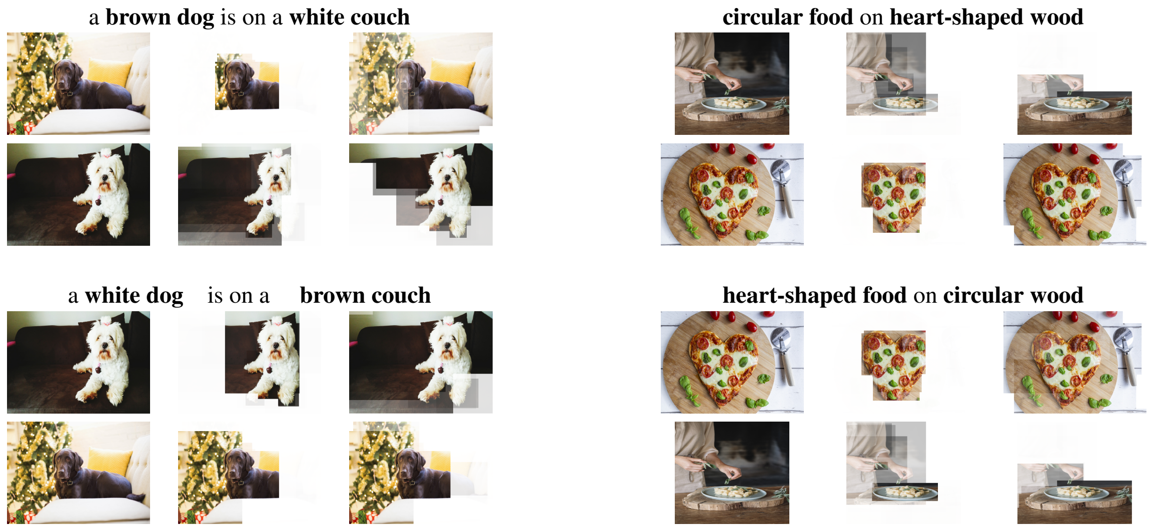


Figure 1. Word-region alignment scores between the image and text features for ViLLA$_{base}$ on examples from Winoground.

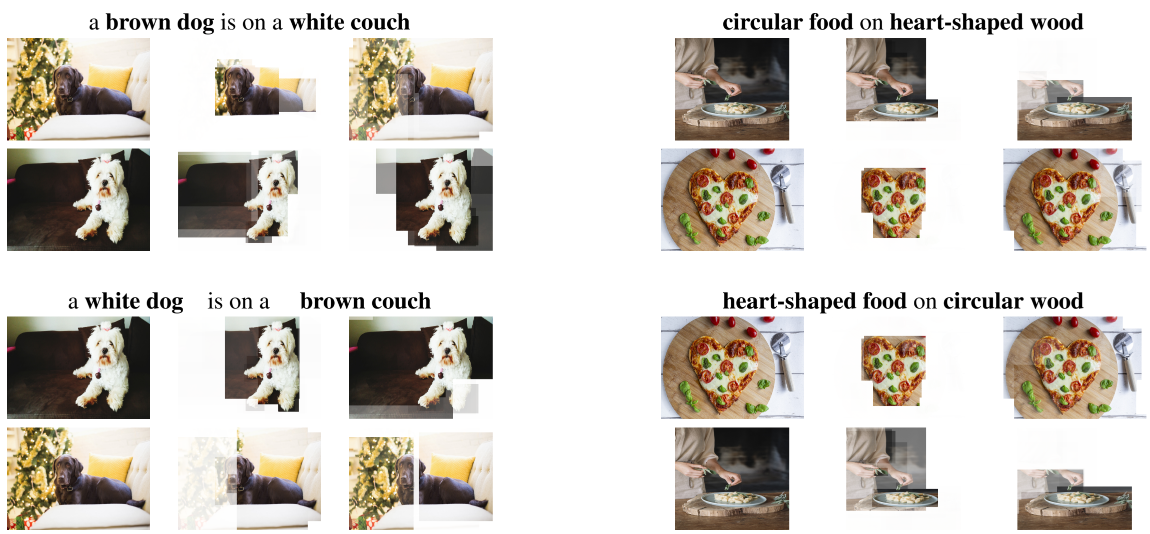


Figure 2. Word-region alignment scores between the image and text features for UNITER$_{base}$ on examples from Winoground.

CVPR
#5457

E. Mechanical Turk Interface

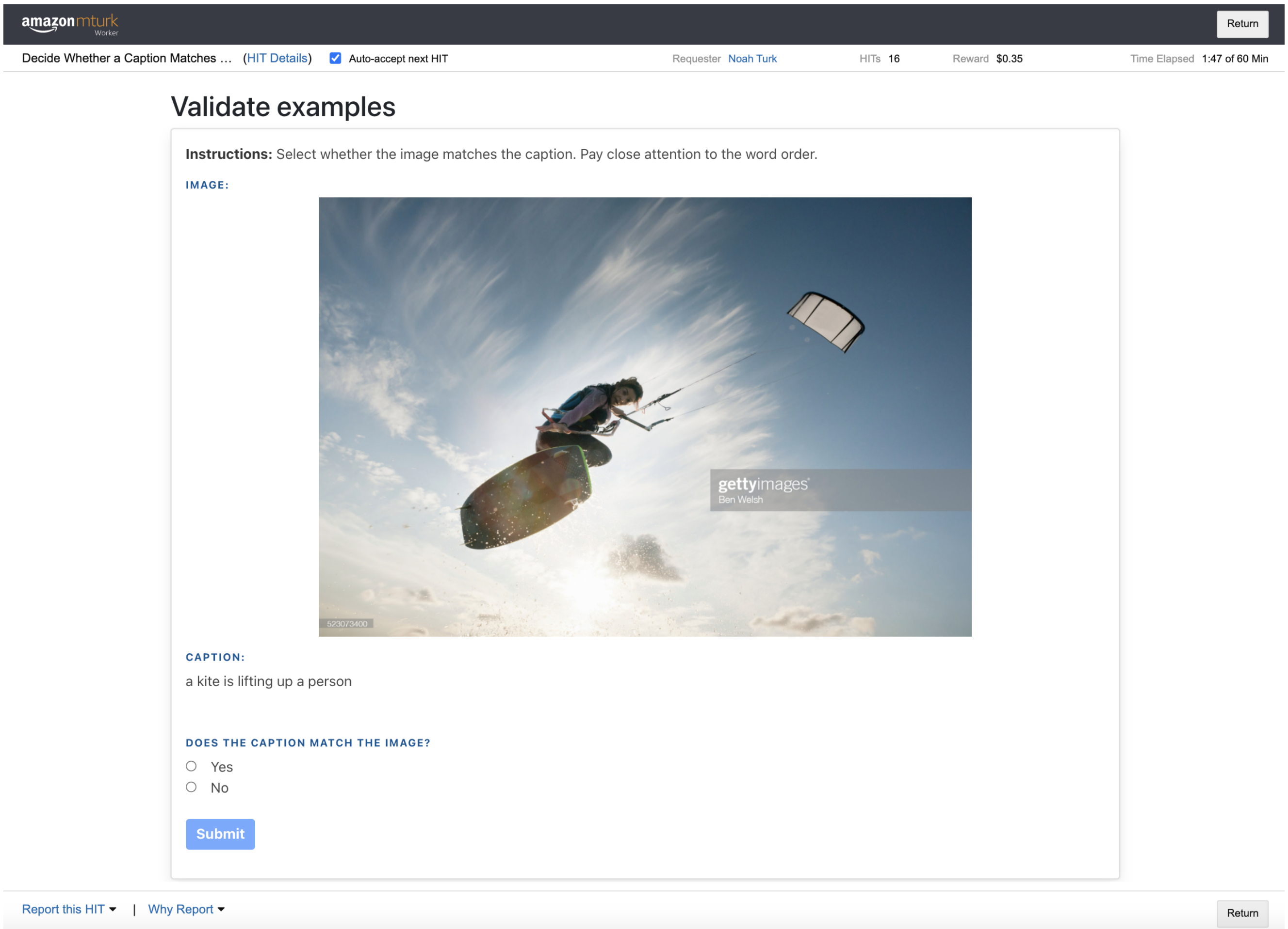


Figure 3. The Amazon Mechanical Turk validation interface. In order to participate, crowdworkers needed to satisfy several criteria: be an English speaker, have 98% previous HIT approval, have completed 1000 previous HITs, and pass the onboarding test. The onboarding test used the same interface as the actual task. It consisted of ten image-caption match questions, with images and captions that are independent from the actual Winoground dataset. If they made one mistake, a pop-up would ask them if they were sure, and they would be allowed to select whether there was a match or not again. If they made any addiitonal mistakes during onboarding, they were disqualified.

F. Ethical Considerations

A key consideration while designing Winoground centered on how the expert annotators would describe the people contained in the images. We avoided using gendered terms (*e.g*. using "person" in place of "woman" or "man") in our captions and did not include any swaps between pairs of captions based on gender, race or ethnicity (*e.g*. "*[the man] hands a water to [the woman]*"). We recognize that, barring direct access to the people in the images, we would be merely making a guess at a person's identity based on our own cultural norms and experiences.

In addition, we encouraged the expert annotators to find images that represent a variety of people across the dimensions of perceived race, gender, disability, *etc*.. We gathered the Getty Images metadata (title and short alt text-like description) and searched them for specific words as a rough proxy for gender representation. The relevant words are either words referring to women (*e.g*. girl, her), words referring to men (*e.g*. boy, him) or words that are gender-neutral (*e.g*. them, themself). Using the Getty Images metadata corresponding to the 800 images in Winoground, 371 images have corresponding metadata that contained at least one word from the lists we created. Using this metadata for these 371 images, we estimate that 152 images only contain women, 123 images only contain men, 22 images only contain people without gender descriptors, and the remaining 74 images contain people described by multiple genders. This serves only as a rough estimate as much of the metadata contain words referring to people that are inherently non-gendered (*e.g*. scuba diver, friend, *etc*.) and because the relevant gendered words we found are themselves subject to the assumptions of those who wrote the titles and captions.