

Appendix

The Appendix is organized as follows:

- Section A includes full illustrations or more experimental results on EQBEN and detailed analysis of EQSIM.
- Section B provides construction details for EQBEN.
- Section C presents the implementation details of EQSIM.
- Section D visualizes more examples in EQBEN.

A. More Results

In this section, we include full illustrations and additional experimental results, due to the space limitation of the main paper.

A.1. Full Ranking Results of Figure 1

In Figure 1 of the main paper, we perform a toy experiment on LAION400M to compare the similarity measure of FIBER and the human oracle. Due to the space limitation, we only show partial ranking results in the main paper. Here we illustrate the full ranking in Figure 7. With the full ranking results, the observation we summarize in the main paper becomes more clear. That is, the similarity changes in FIBER do not faithfully reflect the semantic changes in images (#1 \rightarrow #25) or text queries “right” \rightarrow “left”).

A.2. Retrieval Results on COCO dataset

We report the retrieval performance of FIBER [13] variants on COCO [7] 5K test split in Table 7. We observe similar trends on COCO to that on Flickr30K in Table 2. The results suggest the effectiveness of the proposed EQSIM, which brings large performance gain across all metrics.

Method	Text-to-Image Ret.			Image-to-Text Ret.		
	R@1	R@5	R@10	R@1	R@5	R@10
FIBER [13]	55.19	81.49	88.89	73.39	92.59	96.41
+ FT (COCO) [13]	59.31	83.73	90.43	75.88	93.92	96.79
+ EQSIM	62.55	85.36	91.35	80.16	95.44	97.69

Table 7: Image-text retrieval results on COCO [7] 5K test set. “Ret.” denotes retrieval. Please note that for computation efficiency and fair comparison, we set the image resolution as 288×288 during fine-tuning.

A.3. Full Results of Table 5 and Table 6

We show the full results of ablation studies in Table 8 and Table 9, with group scores across all 5 subsets of EQBEN and Winoground. The observation is similar to the main paper. From Table 8, we can find that EQSIM is scalable in terms of training data, showing the potential to benefit

VL pre-training. The solitary exception happens on EQSD, where EQSIM cannot consistently obtain the improvements. We hypothesize this is probably because EQSD is biased towards the same underlying distribution with the VLMs, as discussed in the main paper. With Table 9, we can find that EQSIM (the hybrid combination of EQSIM_{v1}-all and EQSIM_{v2}-close) is the best-performing one, which validates our claim in Section 3. Meanwhile, EQSIM_{v1}-all and EQSIM_{v2}-close also achieve good results (compared with EQSIM_{v2}-all), where both of them are supported by the claim in Section 3.

A.4. Pilot Study of MLLM on EQBEN

Powered by the remarkable capabilities of the large language model (LLM), the community has witnessed an emergent interest in developing Multimodal Large Language Model (MLLM) [81, 40, 18] very recently. Instead of accepting the pure text as the input, MLLM additionally sees the image and provides the response, which can be regarded as another line of VLMs. Here we conduct a pilot study of the performance of MLLM on our EQBEN. We adopt LLaVa-7B [40] as our base model with Vicuna as the LLM backend. Given two matched image-text pairs $\{I_1, T_1\}$ and $\{I_2, T_2\}$, we concatenate I_1 and I_2 horizontally as the single input image. We build the question prompt with the template: “There are two images (left and right). Now you have two captions: caption 1: $\{T_1\}$; caption 2: $\{T_2\}$. Please indicate which caption corresponds to the left image and which caption corresponds to the right one. The answer should follow the format: “#index for the left image; #index for the right image”. For example, “1;2” represents that caption 1 corresponds to image left.” Since it is hard to reformat the MLLM free-form textual output to the label space, we randomly collect 20 samples from each subset of EQBEN and manually compare the MLLM output and the ground-truth label. The results are shown in Table 10. Interestingly, by comparing two rows, we can find that the performance of MLLM is quite sensitive to the order of the input caption T_1 and T_2 ($\sim 90\%$ v.s $\sim 0\%$). This indicates that the MLLM does NOT truly understand how to distinguish two semantically similar image-text pairs but just follows the given sequence of the captions.

A.5. Computation Cost of EQSIM

We present the computation cost of adding EQSIM in the table below. The forward time is measured with the average of 100 times of forward passes on a single GPU. First, EQSIM is added as a regularization loss, **without** additional overhead on # of parameters. On time cost, we observe an acceptable overhead for fusion-encoder (*i.e.*, METER) due to the similarity calculation on negative pairs. While for dual-encoder (*i.e.*, FIBER), which calculates the similarity for each image-text pair, the extra time needed for EQSIM is



Figure 7: Full ranking results of Figure 1.

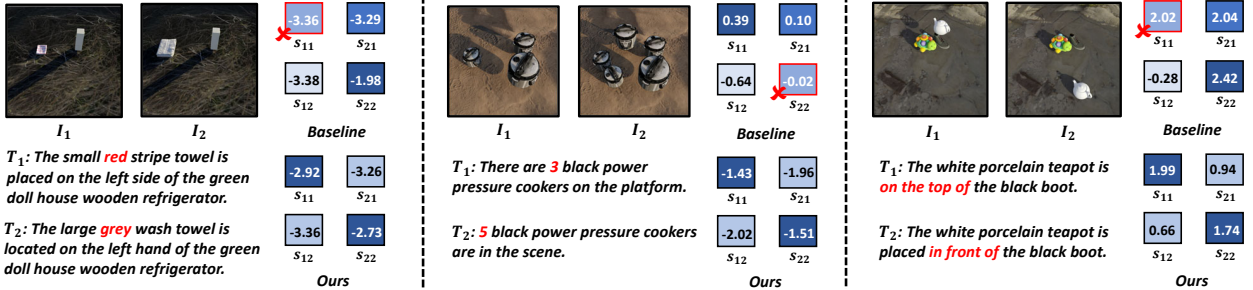
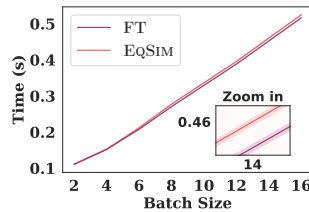


Figure 8: Visualization of similarity scores (s_{ij}) on specific examples of EQBEN. s_{ij} is the similarity score for (I_i, T_j) . Darker color indicates larger similarity. The red cross represents the inferior similarity score leading to the wrong matching result.

almost negligible. Additionally, we show the forward time consumption *v.s.* the batch size in the figure below. The computation cost of EQSIM linearly scales with batch size, which is only slightly higher than the baseline for each data point.

METER		
Method	Time	Parameter
FT	0.384s	319M
EQSIM	0.750s	319M
FIBER		
Method	Time	Parameter
FT	0.270s	242M
EQSIM	0.277s	242M



A.6. More Benchmarking Results on EQBEN

We comprehensively report the model performance of existing VLMs on EQBEN in Table 11. In addition to the observations drawn in the main paper, we can also find that: 1) When comparing the results of ALBEF/BLIP and their variants with contrastive loss (indicated by ‡), utilizing cosine similarity as the similarity measure as in ITC often leads to inferior accuracy compared to score computed by the ITM head. As ITC is usually implemented without cross-attention, making it hard to perform the fine-grained semantic recognition required in EQBEN. 2) Fine-tuning on Flickr30K (F30K) results in a better performance. In contrast to the noisy samples of the pre-training data, F30K contains high-quality captions that describe images in detail, hence helpful for the equivariant similarity learning of VLMs. 3) The recent method BLIP2 shows strong capacity on our EQBEN. Compared to other baselines, it is pre-

trained on a much larger vision-language corpus (with 129 million image-text pairs), and thus shows better generalizability.

A.7. Generalization to Video Grounding

To further validate the generalization ability of the proposed EQSIM, we conduct additional experiments on a very different but relevant downstream task, zero-shot video boundary grounding task [70, 57], where the model is required to accurately predict the video boundary indicating event status change, given the before and after query captions. To adapt a pre-trained VLM to this video-language task, we extract video frames at fps=5 first and measure the similarity between each frame and the two query captions. Then given the two adjacent frames (I_1, I_2) and the two query captions (T_1, T_2) , we define a boundary grounding score $s_{bg} = s(I_1, T_1) + s(I_2, T_2)$ for boundary grounding, where s is the similarity produced by VLMs. s_{bg} actually measures whether the boundary is located between frame I_1 and I_2 . The larger s_{bg} means that (I_1, T_1) and (I_2, T_2) are more likely to be a simultaneous match, thus indicating the boundary between the before and after captions. Results are reported in Table 12 on metrics following [70]. We compute the accuracies based on the absolute distance between ground truth time boundaries and the predicted time boundaries, with the threshold varying from 0.1s to 3s. Across all compared baselines, our EQSIM can attain consistent performance improvements on the average accuracy, suggesting that EQSIM is effective to identify fine-grained shot changes in videos.

FT Data	Method	EQ-AG	EQ-YouCook2	EQ-GEBC	EQ-KUBRIC			EQ-SD	Winoground	Avg
					Location	Counting	Attribute			
F30K [46]	FT	9.24	44.33	8.54	8.95	28.49	66.54	81.29	23.00	33.79
	+ EQSIM	12.64	45.10	10.58	11.05	30.90	70.20	80.70	27.50	36.08
COCO [7]	FT	10.14	42.90	8.93	8.60	26.40	66.60	77.13	21.50	32.77
	+ EQSIM	12.52	45.68	9.37	11.05	26.15	70.75	77.46	25.75	34.84
F30K + COCO	FT	9.96	43.81	8.93	6.94	28.09	64.34	79.90	22.75	33.09
	+ EQSIM	11.98	45.80	10.47	10.89	31.15	70.20	82.62	26.50	36.20
4M [†]	FT	10.49	40.95	7.49	3.25	18.09	65.20	80.63	20.99	30.88
	+ EQSIM	12.78	40.96	9.81	9.20	21.25	67.90	79.77	21.25	32.86

Table 8: Full results of Table 5.

Method	EQ-AG	EQ-YouCook2	EQ-GEBC	EQ-KUBRIC			EQ-SD	Winoground	Avg
				Location	Counting	Attribute			
FT (F30K)	9.24	44.33	8.54	8.95	28.49	66.54	79.97	23.00	33.63
+ HardNeg	12.27	45.43	10.30	10.89	29.49	67.69	81.69	27.00	35.59
+ EQSIM _{v1} -all	12.15	45.71	10.97	9.79	29.94	68.75	81.78	26.49	35.69
+ EQSIM _{v2} -all	12.10	44.97	10.08	10.25	29.05	68.30	79.71	25.49	34.99
+ EQSIM _{v2} -close	13.83	44.99	10.80	11.15	29.25	69.25	80.37	25.75	35.67
+ EQSIM	12.64	45.10	10.58	11.05	30.90	70.20	80.70	27.50	36.08

Table 9: Full results of Table 6.

Caption Order	EQ-AG	EQ-Y.	EQ-G.	EQ-K.	EQ-SD
T_1, T_2	95.00	90.00	90.00	88.33	90.00
T_2, T_1	0.00	0.00	0.00	1.66	0.00

Table 10: Group accuracy (%) of LLaVa on different FT data corpus with different caption input order. Y., G., K. are the short for YouCook2, GEBC and Kubric.

A.8. Distribution Curves on More Subsets

We present the distribution curves of the equivariant score on more EQBEN subsets in Figure 13 as the complement to Figure 6 in the main paper. We can find that our EQSIM (indicated by “Ours”) indeed achieves the most equivariant similarity (*i.e.*, the tightest curve) across different datasets. Meanwhile, it is worth noting that the equivariance of similarity scores are not always positively correlated to the accuracy. For example, on EQ-SD, EQSIM (Ours) is similarly tight as the vanilla fine-tuning (FT), but the accuracy slightly drops.

A.9. More Visualizations

Visualizations of Similarity Scores on Specific Examples. The distribution curves in Figure 6 of the main paper depict the equivariant scores across the whole data. While in Figure 8, we explicitly visualize and compare the similarity scores (blue squares) for specific examples between FIBER baseline and our EQSIM. We can clearly observe that current SoTA VLM still falls short in the similarity measure when facing two visually similar images. On one hand, the matching results are not even correct (red cross); On the other hand, regardless of the correctness of the matching results, the similarity scores are not yet equivariant, similar to the Figure 1(b) of the main paper. As

shown in the left part of Figure 8, for the same visual semantic change (red \leftrightarrow grey), the corresponding similarity change of FIBER $s_{11} - s_{21} = -0.07$ is very different from $s_{22} - s_{12} = 1.4$. While our model can produce much more equivariant similarity measure.

Visualizations of Retrieval Results. We visualize the retrieval results on the commonly used Flickr30K in Figure 9. In addition to the better top-1 retrieval accuracy, our EQSIM can produce much more reasonable similarity measurements for the whole retrieval sequence. For example, for the baseline model, the rank 2 and 3 images are **not** in line with the text of “a young man” and “throw”. While our top-3 images are clearly more relevant.



Figure 9: Visualization of top-5 T2I retrieval results on Flickr30K. Correct (wrong) top-1 images are in green (red).

A.10. EQSIM vs. CyCLIP [17]

After the main paper submission, we notice this related contemporaneous work [17]. We compare and discuss the differences here and will add to the revision.

- Different motivation and implementation. Given the two image-text pairs $\{I_1, T_1\}$ and $\{I_2, T_2\}$, CyCLIP regularizes the CLIP cosine similarity score s with the in-modal consistency (forcing $s(I_1, I_2)$ to be close to $s(T_1, T_2)$) and the cross-modal consistency (forcing $s(I_1, T_2)$ to be close to $s(I_2, T_1)$). While our EQSIM steps from the motivation that the similarity

Method	Natural Subsets									Synthetic Subsets						Avg
	EQ-YouCook2			EQ-GEBC			EQ-AG			EQ-KUBRIC			EQ-SD			
	Text	Image	Group	Text	Image	Group	Text	Image	Group	Text	Image	Group	Text	Image	Group	
LXMERT [60]	13.96	11.98	4.55	13.56	12.73	4.19	18.17	9.02	4.46	18.50	15.35	7.26	11.16	6.15	1.98	10.20
ViLBERT [41]	14.78	12.75	5.18	14.67	12.64	4.82	17.43	8.36	3.89	17.55	18.44	8.13	12.37	7.37	2.78	10.74
CLIP [‡] (RN-50) [48]	47.72	47.99	34.05	10.80	18.03	3.97	14.52	10.44	3.50	21.33	21.93	9.75	90.09	85.92	79.11	33.28
CLIP [‡] (ViT-B/32) [48]	49.48	51.10	36.50	12.57	20.12	4.47	13.91	8.72	3.32	20.56	21.29	9.66	89.16	86.05	78.98	33.73
FLAVA [59]	51.66	54.78	39.68	12.24	16.81	5.07	6.59	13.47	2.15	28.88	28.18	15.90	79.64	84.47	71.10	34.04
ViLT [30]	44.61	46.69	31.74	14.72	16.70	5.62	15.37	9.89	3.45	31.23	27.00	17.90	80.37	79.04	68.93	32.88
ViLT + FT (F30K)	44.06	41.69	29.04	16.43	18.69	7.00	19.00	11.22	5.05	34.00	24.30	16.55	70.12	70.13	55.52	30.85
ALBEF [‡] [35]	51.24	49.09	36.22	9.97	16.04	4.41	10.17	10.92	2.78	18.37	16.98	7.20	82.48	89.95	77.85	32.24
ALBEF	57.01	58.04	44.90	13.56	19.63	5.89	11.28	15.17	3.93	29.87	30.18	18.58	88.96	90.41	83.07	38.03
ALBEF [‡] + FT (F30K)	57.68	54.13	42.74	14.44	19.29	6.28	19.36	12.75	5.41	30.62	22.53	14.28	91.93	92.20	86.51	38.01
ALBEF + FT (F30K)	61.31	61.78	49.28	17.86	22.33	8.32	23.06	16.67	7.56	43.44	36.58	28.23	92.20	91.07	85.32	43.00
BLIP [‡] [34]	55.48	55.60	42.42	13.84	18.68	6.39	13.02	10.46	3.72	27.03	26.03	14.88	82.28	85.26	74.48	35.30
BLIP	59.22	58.36	46.31	15.87	19.79	7.27	19.76	13.87	6.31	29.38	32.25	18.73	85.39	85.52	77.13	38.34
BLIP [‡] + FT (F30K)	58.53	58.61	45.55	15.54	20.62	7.71	18.27	14.33	6.12	34.23	29.95	19.05	91.60	89.82	84.66	39.64
BLIP + FT (F30K)	65.30	64.00	52.96	20.45	23.65	10.08	23.02	18.97	8.03	46.10	38.56	28.93	92.39	92.20	86.31	44.73
BLIP2 [33]	65.78	69.48	55.97	18.25	27.62	9.15	30.46	20.39	11.75	40.24	41.63	28.53	92.59	91.54	86.51	45.19
METER [14]	52.18	49.42	36.81	20.95	18.19	6.95	28.70	15.80	7.88	44.28	35.20	27.26	89.62	84.93	79.44	39.84
+ FT (F30K) [14]	52.68	48.31	36.52	18.08	19.85	7.33	29.50	16.30	8.12	41.11	34.59	24.33	86.64	84.46	77.46	39.02
+ EQSIM	54.12	53.12	40.29	24.20	26.02	11.69	28.85	20.09	10.76	<u>43.68</u>	39.08	28.42	<u>88.04</u>	84.07	<u>77.79</u>	42.28
FIBER [13]	52.04	50.84	38.32	25.19	22.66	11.08	32.49	24.05	13.70	47.94	45.60	33.53	86.05	88.63	79.97	44.86
+ FT (F30K) [13]	57.70	56.46	44.33	18.24	21.33	8.54	26.99	18.69	9.24	50.31	46.06	34.66	90.48	86.64	81.29	43.40
+ EQSIM	58.26	57.10	45.10	<u>21.55</u>	26.07	<u>10.58</u>	<u>29.93</u>	<u>23.42</u>	<u>12.64</u>	51.90	48.40	37.38	90.81	85.98	80.70	45.32

Table 11: Full results on EQBEN. ‡ denotes using cosine similarity of image and text representation as the similarity measure, following the common practice in ITC.

Method		Threshold (s)									Avg
		0.1	0.2	0.5	1.0	1.5	2.0	2.5	3.0		
F30K	FT	2.27	5.18	13.33	26.04	36.16	45.21	53.15	60.01	30.16	
	+ EQSIM	2.93	6.11	15.16	27.06	37.00	46.14	54.07	60.76	31.15	
COCO	FT	2.89	5.81	14.27	26.56	36.52	45.25	53.13	60.19	30.57	
	+ EQSIM	<u>2.82</u>	6.06	15.22	27.18	36.85	45.73	53.68	60.56	31.01	
F30K+COCO	FT	2.65	5.71	13.84	25.60	35.88	44.75	53.04	60.02	30.18	
	+ EQSIM	3.04	6.25	14.89	26.98	36.73	45.53	53.2	60.42	30.88	

Table 12: Comparison between our EQSIM and baselines in video boundary grounding task with regard to different time thresholds.

score change should faithfully respect to the semantic change and derive to the two regularization terms in Eq. (6). Our final objective is the weighted combination of such two terms.

- Different evaluation settings and tasks. CyCLIP is solely built on dual-encoder architecture (e.g. CLIP) and evaluates the effectiveness on the zero-shot image classification task. While our EQSIM can adapt to both dual-encoder and fusion-encoder architectures (e.g. METER and FIBER) and achieve improvements across various VL benchmarks and downstream tasks, e.g., image-text retrieval, vision-language compositionality, and video boundary grounding.
- Better performance of EQSIM. The closest CyCLIP counterpart to our EQSIM is the cross-modal consistency, which we implemented as EQSIM_{v1}-all in Ta-

ble 6. As we compare EQSIM_{v1}-all against +EQSIM, we clearly observe the superior performance of our EQSIM.

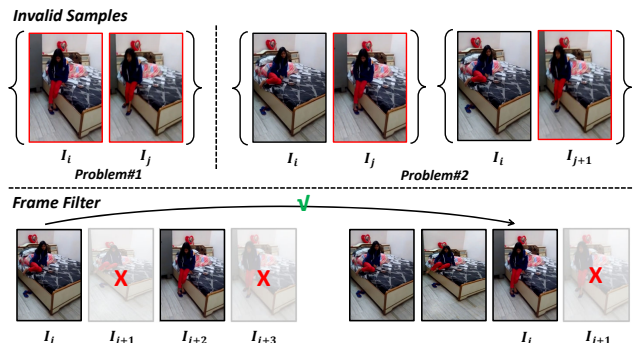


Figure 10: The invalid examples for AG (top) and our proposed frame filter (bottom).

B. Construction Details of EQBEN

In addition to the general construction pipeline of EQBEN in Section 5.3 of the main paper, here we include more details specific to each subset. For all subsets built on natural videos, we denote I_i and I_j as two different frames of the video, while I_{i+1} (I_{j+1}) represents the immediate next frame following I_i (I_j).

B.1. EQ-AG

Source Dataset. Action Genome [26] (AG) captures changes between objects and their pairwise relationships while action occurs. It contains nearly 10K videos with 1.7M visual relationships which can be used for caption generation. Given the scene graph $\langle \text{person} - \text{attention relationship} - \text{spatial relationship} - \text{object} \rangle$, we first create the caption with the template “The person is $\langle \text{attention relationship} \rangle \langle \text{object} \rangle$ which is $\langle \text{spatial relationship} \rangle$ him/her.”

Invalid Samples. In AG, we find that sometimes it is hard to tell apart the two adjacent frames due to the continuity of the video data. This results in two problems for the dataset construction as shown in Figure 10 (top): 1) The two images I_i and I_j are too similar, and may be described by the same caption; 2) The two sample pairs $\{I_i, I_j\}$ and $\{I_i, I_{j+1}\}$ are too similar, leading to many duplicates.

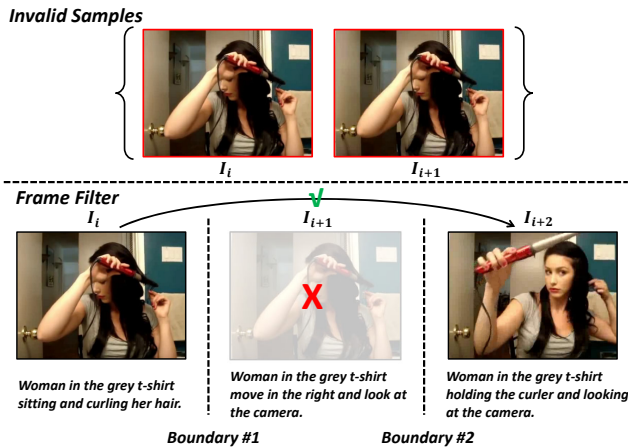


Figure 11: The invalid samples for GEBC (top) and our proposed frame filter (bottom).

Frame Filter. To solve this problem, we adopt a sparse sampling strategy to select frames as candidates (Figure 10 (bottom)). Specifically, we only choose frames I_i and I_j if and only if at least 2 of 3 relationships are different. This will make sure the distinction between two images in a single sample, thus solving problem #1. Furthermore, for problem #2, we assume that given a chosen frame I_i (I_j), the immediate next frame I_{i+1} (I_{j+1}) is too similar to I_i (I_j). Therefore, if I_i (I_j) is chosen, we will skip the subsequent frame (red cross in Figure 10 (bottom)), and move to I_{i+2} (I_{j+2}).

Invalid Samples

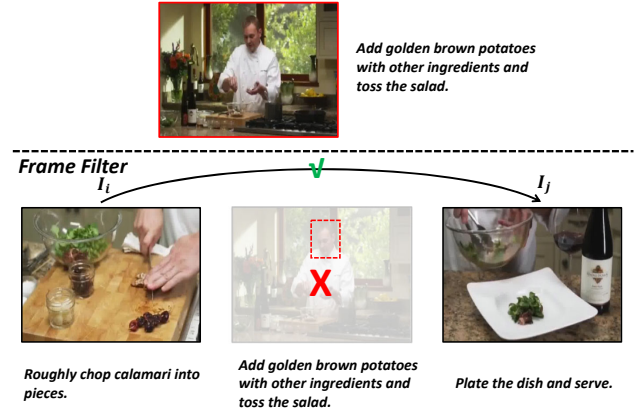


Figure 12: The invalid samples for YouCook2 (top) and our proposed frame filter (bottom).

B.2. EQ-GEBC

GEBC [70] consists of over 170k boundaries associated with captions describing the events before and after the boundaries. It is built upon 12K videos from Kinetic-400 [29] dataset. We construct EQ-GEBC examples based on annotations from the training and validation splits of GEBC. Intuitively, we can directly adopt the frames before and after the boundaries (*i.e.*, I_i and I_{i+1}) as our visual minimally different images, and the provided GEBC annotation before and after the boundaries can be naturally leveraged as the captions.

Invalid Samples. However, similar to AG, we find that it is hard to tell apart the two images separated by a boundary in practice (see Figure 11 (top)). The reason behind is that the boundary of GEBC is annotated as the status change between two video segments (*e.g.*, from “walking” to “running”). Such action words can be hard to recognize from the sampled static frames.

Frame Filter. As shown in Figure 11 (bottom), we propose to skip an additional boundary to choose I_i and I_{i+2} as the twin images to enlarge the semantic gap. Meanwhile, we filter out images with captions containing action words (*e.g.*, “up”, “down”, “upward”, “downward” and “towards”), which are hard to infer without temporal information. Finally, to ensure data quality, we perform a manual screening process with 10 graduate students to filter out invalid samples.

Invalid Samples. As shown in Figure 12 (top), we find that for the cooking video, the chosen frame may contain the view of the chef rather than accurately capturing the objects described in the cooking step. This leads to the mismatch between the image and the caption.

B.3. EQ-YouCook2

We utilize YouCook2 [80] as the data source which contains 2K YouTube videos with average duration of 5.3 min-

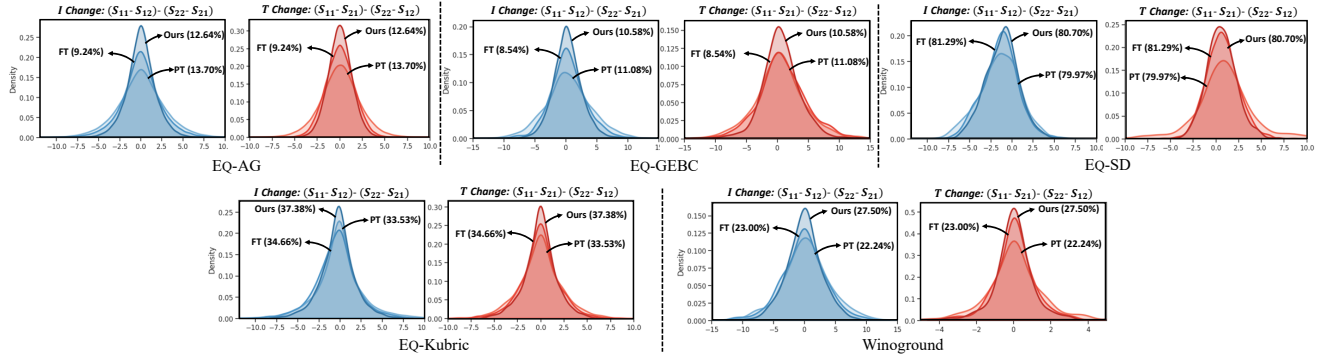


Figure 13: More visualizations of the equivariance score of baselines and our EQSIM based on FIBER [13] on other 4 subsets of EQBEN.

EQ-Kubric	Template	Object / Subject	Verb	Location	Size	Counts	Semantic-Minimally Difference
Location	The $\langle \text{Object} \rangle \langle \text{Verb} \rangle$ $\langle \text{Location} \rangle$ the $\langle \text{Subject} \rangle$.	[green turtle toy, white animal dog toy, black power pressure cooker, great white shark model, white ramekin porcelain, pink damask bath towel, white porcelain teapot, grey elephant toy, bald eagle toy, brown bull, blue gloves, brown hat, white gaming mouse, black keyboard, black boot, white gift box with red straps, red scissors, vintage metal alarm clock, black frypan, stainless steel milk frother, red coffee mug, purple doll house wooden sofa, black sneaker, white canvas shoe, black laptop, black and red gameboy, blue and black backpack, red high heel, stainless steel toaster, white drug bottle, fabric basket, school bus toy, black and yellow hammer, green doll house wooden refrigerator]	[is, is located, is placed]	[in front of, behind, on the left of, on the left side of, on the left hand of, on the right of, on the right side of, on the right hand of, on the top of, above]	N.A.	N.A.	Randomly choose two different locations.
Counting	• There $\langle \text{Verb} \rangle \langle \text{Counts} \rangle$ $\langle \text{Object} \rangle \langle \text{Location} \rangle$. • $\langle \text{Counts} \rangle \langle \text{Object} \rangle$ $\langle \text{Verb} \rangle \langle \text{Location} \rangle$.	• Towel: [blue stripe towel, red stripe towel, green stripe towel, grey wash towel] • Hat: [black hat, grey hat, brown hat] • Plate: [green square saucer plate, red square saucer plate, yellow square saucer plate, green round plate, brown round plate, blue round plate] • Plant container: [blue plant container, yellow plant container, red plant container] • Storage: [metallic mobile device storage, white mobile device storage, blue and white mobile device storage] • Shoe: [brown shoe, green shoe, white running shoe, black shoe, pink shoe, purple running shoe, yellow shoe, blue shoe, red shoe] • Bottle: [yellow drug bottle, red drug bottle, blue drug bottle, purple bottle] • Bowl: [yellow bowl, brown bowl, blue bowl, white bowl, grey dog bowl, pink dog bowl, blue plastic dog bowl]	[is, are]	[in the scene, on the platform]	N.A.	[1, 2, 3, 4, 5]	Randomly choose two different counts.
Attribute	The $\langle \text{Size} \rangle \langle \text{Object} \rangle \langle \text{Verb} \rangle$ $\langle \text{Location} \rangle$ the $\langle \text{Size} \rangle$ $\langle \text{Subject} \rangle$.	• Towel: [blue stripe towel, red stripe towel, green stripe towel, grey wash towel] • Hat: [black hat, grey hat, brown hat] • Plate: [green square saucer plate, red square saucer plate, yellow square saucer plate, green round plate, brown round plate, blue round plate] • Plant container: [blue plant container, yellow plant container, red plant container] • Storage: [metallic mobile device storage, white mobile device storage, blue and white mobile device storage] • Shoe: [brown shoe, green shoe, white running shoe, black shoe, pink shoe, purple running shoe, yellow shoe, blue shoe, red shoe] • Bottle: [yellow drug bottle, red drug bottle, blue drug bottle, purple bottle] • Bowl: [yellow bowl, brown bowl, blue bowl, white bowl, grey dog bowl, pink dog bowl, blue plastic dog bowl]	[is, is located, is placed]	[in front of, behind, on the left of, on the left side of, on the left hand of, on the right of, on the right side of, on the right hand of, on the top of, above]	[small, large]	N.A.	Randomly choose a category set from $\langle \text{Object/Subject} \rangle$; Then randomly choose two different objects with attributes from the set.

Figure 14: Overview of caption generation pipeline for three subsets (*i.e.*, location, counting and attribute) of EQ-KUBRIC. Text in red indicates the aspect of semantic change between two captions.

utes, summing to a total of 176 hours. The videos have been manually annotated with segmentation boundaries and captions. On average there are 7.7 segments/captions per video, and 8.8 words per caption. We construct EQ-YOUCOOK2 examples based on annotations from the training and validation splits of YouCook2. For each video with N segments,

we directly select the middle frame as I_i and its annotated caption as T_i , $i \in \{1, 2, \dots, N\}$.

Frame Filter. To solve this problem, we adopt a simple yet effective solution with the face detector³ for frame filtering. Specifically, we directly discard the frames with human

³https://github.com/ageitgey/face_recognition

EQ-SD	Template	Object	Scene	Attribute	Semantic-Minimally Difference
Object Change	A photo/painting of <i><object></i> <i><scene></i> .	<ul style="list-style-type: none"> [cattle, horse, elephant, goat, deer, camel, zebra] [rabbit, cat, dog, wolf, fox, rat, squirrel] [monkey, koala, panda] [bird, eagle, dove] [duck, chicken] [shark, fish, shrimp, whale, dolphin] [crab, frog] [bear, tiger, lion, pig] 	<ul style="list-style-type: none"> [standing on the grass, in the desert, near the river, in the zoo] [standing on the grass, eating, in the wild, on the desk, on the bench] [standing on the grass, in the jungle, on the tree, in the wild] [in the sky, on the tree, standing on the branch] [on the ground, in the wild] [in the river, under the water] [aside the river, under the water] [in the desert, near the river, in the zoo] 	N.A.	Randomly choose a category set from <i><object></i> . Randomly choose two different objects from the set.
Scene Change	A photo of <i><object></i> (<i><scene></i>).	<ul style="list-style-type: none"> A house on a mountain A river through the valley Modern city street Beach 	<ul style="list-style-type: none"> [at sunset, at night, at winter, at fall, at fog, in the desert] [at sunset with clouds, under noon sunshine, at winter with snow, at fall, at fog] [at sunset with clouds, at winter with snow, at fall, at fog, in heavy raining] [at sunset, at fog, at rainstorm] 	N.A.	Randomly choose a scene from <i><scene></i> .
Attribute Change	A photo/painting of <i><object></i> <i><attribute></i>	[dog, cat]	N.A.	[N.A., wearing a sunglasses, earing a scarf, wearing a crown, wearing a cap, wearing a cowboy hat]	Randomly choose two attributes from <i><attribute></i> .
	A photo/painting of <i><attribute></i> <i><object></i>		N.A.	[black, blue, brown, red]	
	<i><attribute></i> <i><object></i>	[dog, car, bicycle, cat, bus, train, bird, horse, cake, desk]	N.A.	[a real image of, an oil painting of, a pencil sketch of, an Van Gogh post-impressionism painting of]	

Figure 15: Overview of caption generation for EQ-SD dataset. The red color highlights the aspect of semantic change between two captions.

faces.

B.4. EQ-KUBRIC

As introduced in the main paper, EQ-KUBRIC takes advantage of an open-source graphics engine [20] to faithfully generate photo-realistic scene for the given captions. Therefore, the visual-minimal images generation has been translated into the semantic-minimally different captions construction. We categorize the caption change into three aspects: *attribute*, *counting* and *location*. Figure 14 presents the caption construction details. The semantic-minimally difference is ensured by only intervening the corresponding part in the template while leaving other words unchanged.


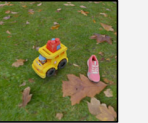
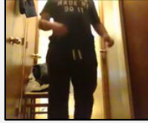
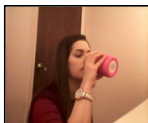
B.5. EQ-SD

Similar process can be applied to EQ-SD, for which we summarize the construction details in Figure 15. We similarly categorize the textual semantic-minimal editing into three aspects: object change, scene change and attribute change. We randomly select from the aforementioned three aspects to construct the semantic-minimally different captions. However, in contrast to the Kubric engine, the generation quality of the stable diffusion model is heavily correlated to the given textual prompt. Therefore, we design a more fine-grained template selection for SD. We select *<scene>* and *<attribute>* from a more restricted subset based on *<object>*. For example, given the object of “horse” and the category of “object change” (first row of Figure 15),

the changed object will be selected from the animals from the same subset (*i.e.*, “cattle”, “elephant”, “goat”, “deer”, “camel” and “zebra”). Meanwhile, the scene shared across two captions will be selected from the first subset of scene (*i.e.*, “standing on the grass”, “in the desert”, “near the river” and “in the zoo”) for rationality.

C. Implementation Details of EQSIM

We fine-tune the models on 8 NVIDIA V100 GPUs. The regularization margin α and the balancing factor β are selected from $\{0, 0.04, 0.1\}$ and $\{0.2, 0.5, 1.0\}$. We adopt an image resolution as 288×288 due to computational constraints. For FIBER, which is implemented with ITC loss for fast retrieval, we adopt the cosine similarity between image and text features as s , and then normalize it by a softmax function. The images and text with top-8 s are regarded as the semantically “close” samples to apply EQSIM_{v2}. METER is designed with ITM loss, which does not compute all pairwise similarities in the training batch. Therefore, we leverage the pre-trained METER model to pre-compute and cache all pairwise similarities in Flickr30K training split, prior to fine-tuning. However, the computation of the ITM similarity for each image-text pair of the training set still takes a long time (more than two weeks on 8 V100 GPUs in practice). To further reduce the computation, we apply a “coarse-to-fine” strategy. For a given image, we first select the top-128 similar images, based on the image feature extracted from the METER vision encoder. As-

EQ-YouCook2	EQ-GEBE	EQ-AG	Attribute	EQ-Kubric Counting	Location	EQ-SD
 Chop a stalk of parsley.	 Baby in white dress sitting on the bed while spreading his hand wide.	 The person is holding the dish which is in front of him/her.	 The large red stripe towel is on the top of the large black keyboard.	 There are 4 school bus toys in the scene.	 The white canvas shoe is placed on the top of the red coffee mug.	 A painting of cat wearing a cap.
 Cut a stick of butter into thin slices.	 Baby in white shirt stop to hold the feet with the hands.	 The person is not contacting the dish which is on the side of him/her.	 The grey wash towel is above the large black keyboard.	 There are 5 school bus toys in the scene.	 The white canvas shoe is located on the right of the red coffee mug.	 A painting of cat wearing a crown.
 Stir the pot.	 Man in the black swim cap and black shorts swimming with his head underwater.	 The person is touching the pillow which is on the side of him/her.	 The yellow bowl is on the top of the small white ramekin porcelain.	 There are 2 red high heels in the scene.	 The blue and black backpack is on the right of the white ramekin porcelain.	 A Van Gogh post-impressionism painting of train.
 Replace the lid on the grill and cook.	 Man in the black swim cap and black shorts lift his head over the surface of water and continues to swim forward.	 The person is carrying the pillow which is behind him/her.	 The large brown bowl is above the small white ramekin porcelain.	 4 red high heels are in the scene.	 The blue and black backpack is located on the top of the white ramekin porcelain.	 A pencil sketch of train.
 Trim the fat off some beef and place in a slow cooker.	 Man in yellow t-shirt and blue athletic suit holding the barbell on shoulders in front.	 The person is holding the clothes which is in front of him/her.	 The large pink shoe is on the right of the large school bus toy.	 2 white ramekin porcelains are on the platform.	 The white drug bottle is on the right of the white animal dog toy.	 A painting of squirrel eating.
 Sprinkle the spices over the beef cubes and stir.	 Man in yellow t-shirt and blue athletic suit take back one leg to stand straight.	 The person is not contacting the clothes which is behind him/her.	 The red shoe is placed on the right of the large school bus toy.	 5 white ramekin porcelains are on the platform.	 The white drug bottle is on the left side of the white animal dog toy.	 A painting of fox eating.
 Cook the mushrooms in a pan.	 Man in blue tank t-shirt and white shorts running on the track.	 The person is drinking from the bottle which is in front of him/her.	 The large brown hat is placed in front of the white porcelain teapot.	 There is 1 red coffee mug in the scene.	 The blue and black backpack is located on the right of the brown bull.	 A photo of a blue desk.
 Cook the carrots in a pan.	 Man in blue tank t-shirt and white shorts jump and roll body in the air.	 The person is holding the bottle which is on the side of him/her.	 The large black hat is in front of the white porcelain teapot.	 There are 4 red coffee mugs in the scene.	 The blue and black backpack is placed in front of the brown bull.	 A photo of a brown desk.

Natural

Synthetic

Figure 16: More visualizations of the examples for our EQBEN.

suming each image is associated with 5 captions, we then utilize the ITM head to compute a fine-grained similarity measure for 128×5 image-text pairs (leading to 1 hour on 8 V100 GPUs). During retrieval fine-tuning, we follow the original METER to sample 15 captions as negatives and additionally sample their counterpart images for EQSIM. Furthermore, 8 of 15 items (*i.e.*, $k = 8$) are selected as hard negative (*i.e.*, semantically close) samples based on the pre-computed similarity matrix. In METER, the similarity score s is normalized with a sigmoid activation. We apply other model-specific hyper-parameters (*e.g.*, training epochs and learning rates) following the original METER [14] and FIBER paper [13].

D. More Examples of EQBEN

Figure 16 visualizes examples in EQBEN. We can clearly find that the two images from one data sample are visually similar, indicating that our EQBEN indeed focuses on visual-minimal change.

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint*, 2022. [2](#)
- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015. [3](#)
- [3] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992. [4](#)
- [4] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021. [3](#)
- [5] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *NeurIPS*, 32, 2019. [3](#)
- [6] Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. Attacking visual language grounding with adversarial examples: A case study on neural image captioning. *arXiv preprint arXiv:1712.02051*, 2017. [3](#)
- [7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO Captions: Data collection and evaluation server. *arXiv preprint*, 2015. [3](#), [5](#), [8](#), [9](#), [11](#)
- [8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *ECCV*, 2020. [2](#)
- [9] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*, 2022. [1](#)
- [10] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *ICML*, pages 2990–2999. PMLR, 2016. [3](#)
- [11] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *arXiv preprint arXiv:1801.10130*, 2018. [3](#)
- [12] Rumén Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljačić. Equivariant contrastive learning. *arXiv preprint arXiv:2111.00899*, 2021. [3](#)
- [13] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *arXiv preprint arXiv:2206.07643*, 2022. [1](#), [2](#), [6](#), [7](#), [8](#), [9](#), [12](#), [14](#), [17](#)
- [14] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. In *CVPR*, 2022. [1](#), [2](#), [6](#), [7](#), [12](#), [17](#)
- [15] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020. [2](#)
- [16] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. Vision-language pre-training: Basics, recent advances, and future trends. *arXiv preprint arXiv:2210.09263*, 2022. [2](#)
- [17] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *arXiv preprint arXiv:2205.14459*, 2022. [3](#), [11](#)
- [18] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023. [9](#)
- [19] Jonathan Gordon, David Lopez-Paz, Marco Baroni, and Diane Bouchacourt. Permutation equivariant models for compositional generalization in language. In *ICLR*, 2019. [3](#)
- [20] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanaprasgam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: a scalable dataset generator. 2022. [4](#), [5](#), [15](#)
- [21] Tanmay Gupta, Ryan Marten, Aniruddha Kembhavi, and Derek Hoiem. Grit: General robust image task benchmark. *arXiv preprint arXiv:2204.13653*, 2022. [3](#)
- [22] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33:5679–5690, 2020. [3](#)
- [23] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021. [3](#)
- [24] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. [4](#), [5](#)
- [25] Hexiang Hu, Ishan Misra, and Laurens van der Maaten. Evaluating text-to-image matching using binary image selection (bison). In *ICCV Workshops*, pages 0–0, 2019. [3](#)
- [26] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *CVPR*, pages 10236–10247, 2020. [4](#), [5](#), [13](#)
- [27] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. [1](#), [2](#)
- [28] Carlos E Jimenez, Olga Russakovsky, and Karthik Narasimhan. Carets: A consistency and robustness evalu-

- ative test suite for vqa. *arXiv preprint arXiv:2203.07613*, 2022. 3
- [29] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 13
- [30] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, pages 5583–5594. PMLR, 2021. 7, 12
- [31] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 8
- [32] Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. Image retrieval from contextual descriptions. In *ACL*, pages 3426–3440, 2022. 3
- [33] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 12
- [34] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint*, 2022. 1, 2, 7, 12
- [35] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 1, 2, 7, 12
- [36] Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *ICCV*, pages 2042–2051, 2021. 3
- [37] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint*, 2019. 2
- [38] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantic aligned pre-training for vision-language tasks. In *ECCV*, 2020. 2
- [39] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 4
- [40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 9
- [41] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 2, 7, 12
- [42] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *CVPR*, pages 12700–12710, 2021. 3
- [43] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2Text: Describing images using 1 million captioned photographs. In *NeurIPS*, 2011. 8
- [44] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*, 2021. 1, 2, 3, 5, 6, 7
- [45] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020. 3
- [46] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015. 2, 3, 4, 5, 6, 8, 11
- [47] Guo-Jun Qi, Liheng Zhang, Feng Lin, and Xiao Wang. Learning generalized transformation equivariant representations via autoencoding transformations. *TPAMI*, 2020. 3
- [48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 7, 12
- [49] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 5
- [50] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *TPAMI*, 2016. 2, 7
- [51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 4, 7
- [52] Halsey Lawrence Royden and Patrick Fitzpatrick. *Real analysis*, volume 32. Macmillan New York, 1988. 2
- [53] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 5
- [54] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1
- [55] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 8
- [56] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sanginetto, and Raffaella Bernardi. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*, 2017. 3
- [57] Mike Zheng Shou, Stan Weixian Lei, Weiyao Wang, Deepti Ghadiyaram, and Matt Feiszli. Generic event boundary de-

- tection: A benchmark for event segmentation. In *ICCV*, pages 8075–8084, 2021. 10
- [58] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *CVPR*, 2022. 2, 7
- [59] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, pages 15638–15650, 2022. 7, 12
- [60] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 1, 2, 7, 12
- [61] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, pages 3716–3725, 2020. 3
- [62] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visiolinguistic compositionality. In *CVPR*, pages 5238–5248, 2022. 1, 2, 3, 4, 5, 6, 7, 8
- [63] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999. 4
- [64] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015. 3
- [65] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint*, 2022. 2
- [66] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *CVPR*, pages 10760–10770, 2020. 3
- [67] Tan Wang, Zhongqi Yue, Jianqiang Huang, Qianru Sun, and Hanwang Zhang. Self-supervised learning disentangled group representation as feature. *NeurIPS*, 34:18225–18240, 2021. 3
- [68] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 1
- [69] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint*, 2021. 2
- [70] Yuxuan Wang, Difei Gao, Licheng Yu, Weixian Lei, Matt Feiszli, and Mike Zheng Shou. Geb+: A benchmark for generic event boundary captioning, grounding and retrieval. In *ECCV*, pages 709–725. Springer, 2022. 4, 5, 10, 13
- [71] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. In *ICLR*, 2022. 2
- [72] Maurice Weiler and Gabriele Cesa. General e (2)-equivariant steerable cnns. *NeurIPS*, 32, 2019. 3
- [73] Yuyang Xie, Jianhong Wen, Kin Wai Lau, Yasar Abbas Ur Rehman, and Jiajun Shen. What should be equivariant in self-supervised learning. In *CVPR*, pages 4111–4120, 2022. 3
- [74] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mostafa Seyedhosseini, and Yonghui Wu. CoCa: Contrastive captioners are image-text foundation models. *arXiv preprint*, 2022. 1, 2
- [75] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Guntjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 5
- [76] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *ECCV*, 2016. 3
- [77] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6720–6731, 2019. 3
- [78] Pengchuan Zhang, Xiujuan Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. VinVL: Revisiting visual representations in vision-language models. In *CVPR*, 2021. 2
- [79] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. VI-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022. 3
- [80] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 4, 5, 13
- [81] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 9