1 Appendix

We provide additional details for datasets, experimental settings, results, and analysis in the supplementary material.

A. Dataset details

Pre-training datasets. Instead of using well-curated datasets, we use image-AltText pairs sampled from a web-crawled dataset [42]. We collect 300M image-text pairs from the Web and denote it as WIT-300M. Based on WIT-300M, we build four subsets to cover from small to large scales. Specifically, WIT-200M is a subset of WIT-300M. WIT-100M is a subset of WIT-200M. WIT-12M is a subset of WIT-100M. WIT-3M is a subset of WIT-12M.

Table A1: Details of 9 VTAB zero-shot classification datasets.

Dataset	Metric	Categories	Train Size	Test Size
CIFAR-100 [19]	Accuracy	100	50,000	10,000
SVHN [46]	Accuracy	10	73,257	26,032
DTD [10]	Accuracy	47	3,760	1,880
Oxford Pets [30]	Mean per class	37	3,680	3,669
Caltech101 [15]	Mean per class	102	3,060	6,085
Flowers102 [29]	Mean per class	102	2,040	6,149
EuroSAT [17]	Accuracy	10	10,000	5,000
RESISC45 [8]	Accuracy	45	25,200	6,300
Camelyon [40]	Accuracy	2	262,144	32,768

VTAB datasets. We choose 9 classification datasets suitable for zero-shot evaluation from VTAB [47]. Table A1 summarizes zero-shot image classification datasets. For both original CLIP models and our models, we use the identical prompt set from CLIP. Every class label is expanded using a collection of prompt templates, as defined by CLIP, including examples like "A photo of a [classname]." The class embedding is then computed by taking the average of the embeddings of all such templates, followed by L2-normalization.

B. Implementation details

Pre-training hyper-parameters. We summarize the pre-training hyper-parameters for CLIP training in Table A2. We pre-train models on up to 512 TPUs with JAX [3].

Table A2: Details of the pre-training hyper-parameters for CLIP training on our web-crawled datasets.

Config	Value
Batch size	$8,192$
Optimizer	AdamW
Learning rate	5×10^{-4}
Weight decay	0.5
Adam β	$\beta_1, \beta_2 = (0.9, 0.98)$
Adam ϵ	1×10^{-8}
Total epochs	40
Warm up epochs	1
Learning rate schedule	cosine decay

(a) Pre-training hyper-parameters on 3M.

Config	Value
Batch size	$8,192$
Optimizer	AdamW
Learning rate	5×10^{-4}
Weight decay	0.5
Adam β	$\beta_1, \beta_2 = (0.9, 0.98)$
Adam ϵ	1×10^{-8}
Total epochs	35
Warm up epochs	1
Learning rate schedule	cosine decay

(b) Pre-training hyper-parameters on 12M.

Config	Value
Batch size	$32,768$
Optimizer	AdamW
Learning rate	5×10^{-4}
Weight decay	0.2
Adam β	$\beta_1, \beta_2 = (0.9, 0.98)$
Adam ϵ	1×10^{-6}
Total epochs	32
Warm up iterations	$2,000$
Learning rate schedule	cosine decay

(c) Pre-training hyper-parameters on 100M

Config	Value
Batch size	$32,768$
Optimizer	AdamW
Learning rate	5×10^{-4}
Weight decay	0.2
Adam β	$\beta_1, \beta_2 = (0.9, 0.98)$
Adam ϵ	1×10^{-6}
Total epochs	32
Warm up iterations	$2,000$
Learning rate schedule	cosine decay

(d) Pre-training hyper-parameters on 200M.

C. More experimental results

In this section, we present more detailed experimental results and our ablation studies (e.g., generalization of VeCLIP with a large backbone, public and well-curated datasets for pre-training).

C.1. Larger backbone architectures

We also investigate the performance of VeCLIP using a larger backbone architecture, ViT-L/14. The comparison results are summarized in Table A3. First, VeCLIP shows a consistent improvement over CLIP employing ViT-L/14 across all downstream tasks. Second, VeCLIP utilizing ViT-L/14 surpasses its counterpart employing ViT-B/16, notably excelling in image classification tasks, achieving a notable improvement of over 5% on both ImageNet and ImageNetV2. This shows that VeCLIP has the potential to be scalable with larger backbone architectures and larger-scale datasets.

Table A3: Ablation studies on different backbones with VeCLIP. We use 200M as the pre-training dataset.

Model	Backbone	COCO (R@1)		Flickr30k (R@1)		ImageNet	ImageNetV2
		I2T	T2I	I2T	T2I		
CLIP	ViT-B/16	52.20	34.97	80.90	63.23	63.72	56.84
VeCLIP	ViT-B/16	**67.20**	**48.40**	**91.10**	**76.32**	**64.62**	**57.67**
Performance Gain		**+15.00**	**+13.43**	**+10.20**	**+13.06**	**+0.90**	**+0.81**
CLIP	ViT-L/14	53.92	37.86	84.60	66.78	68.51	61.13
VeCLIP	ViT-L/14	**69.92**	**51.32**	**92.60**	**79.04**	**69.85**	**63.54**
Performance Gain		**+16.00**	**+13.46**	**+8.00**	**+12.26**	**+1.34**	**+2.41**
VeCLIP ViT-L/14 vs B/16		**+2.72**	**+2.92**	**+1.50**	**+2.72**	**+5.23**	**+5.87**

Table A4: Ablation studies on well-curated datasets (CC3M and CC12M [5]) and the effect of data quality with ViT-B/16 as the vision backbone.

Model	Model	COCO (R@1)		Flickr30k (R@1)		ImageNet	ImageNetV2
		I2T	T2I	I2T	T2I		
WIT-3M	CLIP	5.18	3.40	10.50	6.88	8.02	6.88
	VeCLIP	**22.30**	**13.01**	**40.60**	**27.58**	**15.98**	**13.51**
Performance Gain		**+17.12**	**+9.61**	**+30.10**	**+20.70**	**+7.96**	**+6.63**
CC3M	CLIP	13.88	9.64	26.30	18.04	14.59	12.52
	VeCLIP	**32.04**	**22.07**	**57.20**	**36.54**	**20.73**	**17.90**
Performance Gain		**+18.16**	**+12.43**	**+30.90**	**+18.50**	**+6.14**	**+5.38**
WIT-12M	CLIP	22.58	14.23	44.40	30.90	31.14	25.91
	VeCLIP	**47.78**	**31.62**	**73.90**	**55.68**	**38.11**	**32.51**
Performance Gain		**+25.20**	**+17.39**	**+29.50**	**+24.78**	**+6.97**	**+6.60**
CC12M	CLIP	37.96	24.40	59.70	44.90	39.24	34.41
	VeCLIP	**53.23**	**36.90**	**75.20**	**62.10**	**45.32**	**40.21**
Performance Gain		**+15.27**	**+12.50**	**+15.50**	**+17.20**	**+6.08**	**+5.80**

C.2. Generalization on well-curated datasets: CC3M and CC12M

Besides our crawled noisy WIT datasets, we also use a well-curated dataset, e.g., CC3M and CC12M [5], to show the effectiveness and generalizability of our proposed approach on well-curated datasets. CC3M and CC12M [5] were curated via several rounds of comprehensive refining and filtering to get high-quality image-caption pairs. We show high-quality examples of CC3M and the comparison of CC3M's captions and WIT-3M's AltTexts in Appendix D. We present an experimental comparison between our crawled WIT datasets and well-curated CC3M/CC12M [5] in this subsection.

3M. As shown in Table A4, CC3M outperforms WIT-3M when coupled with CLIP pre-training, yielding a notable increase of +10.70% on the COCO I2T task. Additionally, VeCLIP exhibits substantial improvement for both WIT-3M and CC3M. Notably, we achieve a remarkable over 30% improvement on the I2T task in Flickr30K, and an impressive over 5% boost on ImageNet and ImageNetV2.

12M. Similar to 3M settings, CC12M exhibits superior quality and attains better results in contrast to WIT-12M when utilized with CLIP and original AltTexts. VeCLIP demonstrates notable improvements for both WIT-12M and CC12M. For instance, VeCLIP yields a remarkable +12.27% increase in the I2T task of COCO, along with an impressive over 5% improvement on both ImageNet

and ImageNetV2. These findings emphasize the effectiveness and generalizability of VeCLIP in both noisy web-crawled datasets and meticulously curated datasets, where a richer set of visual concepts is harnessed for pre-training.

C.3. Complete visual descriptions vs simplified entity representations

In Table 6b of the main paper, we note that sole training on VeCap might detriment zero-shot performance in comparison to the original AltText. Conversely, our mixed training approach yields optimal outcomes. This intriguing finding propels us toward a more profound investigation of zero-shot classification tasks. Following established works [13, 33], we employ an identical set of prompting templates, such as "a photo of a [CLS]" for ImageNet [11]. It is conceivable that this direct and uncomplicated prompt may diverge significantly from VeCap's pre-training, which encompasses a more extensive and intricate set of visual concepts. To address this, we reformulate VeCap into a format as Simplified Entity Representation (SER). Specifically, we employ the NLTK package to extract entities from VeCap and subsequently apply filtering to retain only noun entities, denoted as $(A, B, C...) \in U$. This transformation results in VeCap being presented as "a photo of [U]", offering a concise representation of all extracted entities. The results are summarized in Table A5. Surprisingly, we find that even with SER-style captions, the zero-shot performance remains inferior to that achieved with the original AltText. We hypothesize that this discrepancy may arise from a lack of data diversity. When all sentences adhere to the same distribution, there exists a risk of overfitting in the pre-trained model, resulting in suboptimal performance in downstream tasks.

Table A5: Ablation studies on VeCap and Simplied Entities Representation (SER). We use ViT-B/16 as the backbone and use 200M as the pre-trained dataset.

Model	Caption	COCO (R@1)		Flickr30k (R@1)		ImageNet	ImageNetV2
		I2T	T2I	I2T	T2I		
CLIP	AltText	52.20	34.97	80.90	63.23	63.72	56.84
VeCLIP	SER	65.88	**49.04**	89.20	75.96	58.58	52.89
VeCLIP	VeCap	**67.20**	48.40	**91.10**	**76.32**	**64.62**	**57.67**

C.4. Complementary to DFN

In Section 4.4, we discuss VeCap is complemtentary to DFN [14]. We provide another example where we use B-16 as the backbone and set the image resolution as 224: as shown in Table A6, our VeCap is complementary to DFN [14].

C.5. Main results with WIT-300M

We show the detailed results with the Web-crawled Image-Text 300M dataset (WIT-300M) here. We summarize the results on various downstream tasks in

Table A6: VeCap is complementary to DFN.

Data	COCO (I2T)	COCO (T2I)	ImageNet
DFN	62.96	43.20	76.15
DFN + VeCap	66.28	45.12	76.19
Gain	**+3.32%**	**+1.92%**	**+0.04%**

Table A9. There are two major observations. First, we observe that the results obtained with a dataset size of 300M are close to those achieved with 200M for both CLIP and VeCLIP models. This suggests that a dataset scale of 200 million is sufficient for effectively training a ViT-B/16-based CLIP model. Second, VeCLIP achieves significant improvement on retrieval tasks even under 300M settings. Nevertheless, the improvement observed in ImageNet/ImageNetV2 is marginal.

Table A7: Zero-shot classification accuracy. Top-1 Accuracies (%) of VTAB [47] across 9 tasks (6 from natural and 3 from specialized sets) are reported.

Data	Model	Natural Sets						Specialized Sets			Average
		Caltech101	CIFAR100	SVHN	DTD	OxPet	Flowers102	EuroSAT	RESISC45	Camelyon	
Model Architecture: ViT-B/16											
3M	CLIP	39.50	9.83	**20.89**	7.42	7.44	10.40	**11.94**	7.93	50.65	18.45
	VeCLIP	**54.30**	**17.74**	18.74	**11.23**	**10.09**	**22.75**	7.35	**16.54**	**52.52**	**23.48**
Performance Gain		**+14.80**	**+7.91**	**-2.15**	**+3.81**	**+2.65**	**+12.35**	**-4.59**	**+8.61**	**+1.87**	**+5.03**
12M	CLIP	70.43	30.06	**30.11**	30.69	34.51	33.67	8.87	30.05	53.46	35.76
	VeCLIP	**70.58**	**45.10**	23.61	**30.90**	**36.22**	**43.94**	**27.46**	**38.09**	**55.54**	**41.27**
Performance Gain		**+0.15**	**+15.04**	**-6.50**	**+0.21**	**+1.71**	**+10.27**	**+18.59**	**+8.04**	**+2.08**	**+5.51**
100M	CLIP	81.44	54.75	38.70	57.28	**70.51**	51.71	34.45	48.56	53.87	54.59
	VeCLIP	**81.64**	**64.62**	**46.49**	**57.51**	64.81	**66.41**	**46.23**	**51.75**	**58.51**	**59.78**
Performance Gain		**+0.20**	**+9.87**	**+7.79**	**+0.23**	**-5.70**	**+14.70**	**+11.78**	**+3.19**	**+4.64**	**+5.19**
200M	CLIP	82.30	61.87	42.83	**64.29**	**75.60**	58.67	46.73	**55.59**	59.30	60.79
	VeCLIP	**83.14**	**68.14**	**44.93**	61.95	72.61	**68.51**	**47.36**	55.10	**62.59**	**62.70**
Performance Gain		**+0.84**	**+6.27**	**+2.10**	**-2.34**	**-2.99**	**+9.84**	**+0.63**	**-0.49**	**+3.29**	**+1.91**
300M	CLIP	**83.58**	63.36	50.04	**66.16**	74.30	61.81	39.95	56.44	**53.94**	61.06
	VeCLIP	83.07	**68.37**	**50.07**	65.98	**75.36**	**69.71**	**48.28**	**58.09**	51.94	**63.43**
Performance Gain		**-0.51**	**+5.01**	**+0.03**	**-0.18**	**1.06**	**+7.90**	**+8.33**	**+1.65**	**-2.00**	**+2.37**

As shown in Table A10, our VeCLIP with DFN [14] can outperform FLIP [23] and OpenAI CLIP with different backbones. Specifically, our ViT-H/14 model achieves impressive 83.1% of accuracy on ImageNet. We leave the further study of combing the synthetic data (VeCap) with other data curation approaches as a future work.

D. Performance trend across scales

Besides the performance gain, we also visualize the performance trend across data scales in pre-training. As shown in A1, the performance of CLIP utilizing original AltTexts exhibits a marked surge with the increased data size: while its

Table A8: Image-to-Image retrieval results (mAP) on 6-domain GPR1200.

Data	Model	Domain Name						All
		Land	Faces	iNat	INST	Sketch	SOP	
3M	CLIP	57.98	20.76	17.61	31.14	18.23	74.29	36.67
	VeCLIP	**66.55**	**23.51**	**20.43**	**38.63**	**24.59**	**77.65**	**41.89**
12M	CLIP	74.47	30.65	23.60	52.15	30.68	84.25	49.30
	VeCLIP	**79.30**	**31.72**	**25.53**	**56.65**	**41.42**	**84.69**	**53.22**
100M	CLIP	**85.64**	**51.68**	29.66	68.19	42.45	90.38	61.33
	VeCLIP	85.59	42.83	**30.72**	**71.96**	**52.59**	**90.54**	**62.37**
200M	CLIP	**86.96**	**56.54**	30.95	71.51	46.03	90.95	63.83
	VeCLIP	86.40	48.48	**31.72**	**73.74**	**56.52**	**91.16**	**65.67**
300M	CLIP	**87.17**	**57.09**	31.83	72.80	47.03	**91.30**	64.54
	VeCLIP	86.22	48.51	**32.05**	**75.29**	**56.18**	91.25	**66.91**

Table A9: Zero-shot classification results (Top-k Accuracy) on ImageNet and ImageNetV2.

Data	Model	ImageNet			ImageNetV2		
		Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
3M	CLIP	5.46	21.05	28.70	7.09	18.52	25.83
	VeCLIP	**15.98**	**34.11**	**43.23**	**13.51**	**30.03**	**38.93**
12M	CLIP	31.60	58.80	69.49	27.03	52.68	63.37
	VeCLIP	**38.11**	**66.74**	**76.36**	**32.53**	**60.16**	**70.50**
100M	CLIP	58.64	85.82	91.79	50.96	79.77	86.91
	VeCLIP	**60.77**	**87.77**	**93.16**	**54.17**	**82.51**	**89.24**
200M	CLIP	63.72	89.26	94.11	56.84	83.50	89.79
	VeCLIP	**64.62**	**90.27**	**94.90**	**57.67**	**85.24**	**91.62**
300M	CLIP	65.70	90.55	94.87	58.58	85.32	91.35
	VeCLIP	**65.71**	**91.15**	**95.36**	**58.76**	**86.31**	**91.95**

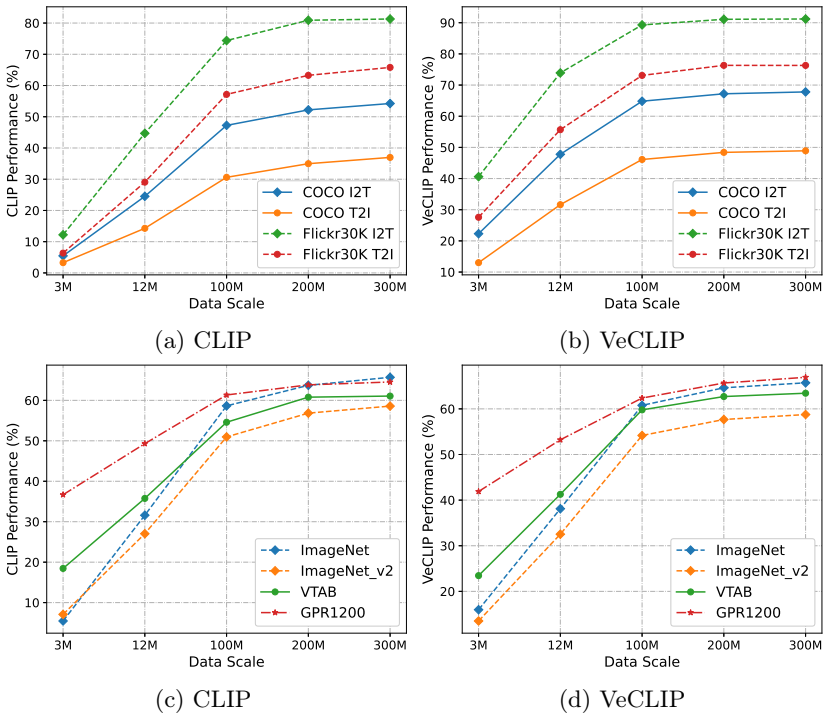
Table A10: Comparison bwetween VeCLIP and other models.

Backbone	Model	Data	COCO (R@1)		Flickr30k (R@1)		ImageNet
			I2T	T2I	I2T	T2I	
ViT-B/16	OpenAI CLIP	OpenAI-400M	53.8	33.1	88.0	68.7	68.6
	FLIP [23]	LAION-400M	-	-	-	-	68.0
	VeCLIP	DFN [14] + VeCap	**66.3**	**45.1**	**88.8**	**73.6**	**76.2**
ViT-L/14	OpenAI CLIP	OpenAI-400M	58.4	37.8	88.0	68.7	75.3
	FLIP [23]	LAION-400M	60.2	44.2	89.1	75.4	74.6
	VeCLIP	DFN [14] + VeCap	**71.1**	**51.1**	**93.1**	**81.0**	**82.0**
ViT-H/14	VeCLIP	DFN [14] + VeCap	**72.8**	**52.3**	**93.6**	**82.6**	**83.1**

starting point is poor at 3M, it demonstrates swift progression up to 12M and 100M. However, once scaled beyond 100 million, the performance trend exhibits a gradual and eventually saturated growth. On the other hand, commencing with a higher baseline, VeCLIP employing VeCap demonstrates substantial improvement in comparison to CLIP within small to medium scales (3M and 12M). As we progress beyond 300M, the performance gains of VeCLIP become relatively incremental but still noticeable in retrieval tasks. Both CLIP and VeCLIP reach a saturation point when scaled up to 100M: once over 100M, the performance gain becomes less impressive as a relatively small but high-quality dataset may reach the upper bound of the learning ability constrained by the model's architecture.

E. Caption quality comparison between well-curated Datasets and WIT datasets

In Appendix C.2, we find CLIP performs notably better when pre-trained on CC3M compared to the case of being pre-trained on noisy crawled WIT datasets due to several rounds of filtering and refining involved in the curation of CC3M and CC12M. In this section, we show more data analysis in terms of the number of unique verbs, adjectives, nouns, and average token lengths. We randomly sample 10K data and summarize the results in Table A11.



(a) CLIP (b) VeCLIP

(c) CLIP (d) VeCLIP

Fig. A1: Performance trend with ViT-B/16 as the vision backbone. (a) and (c) show the trend of CLIP with original AltTexts while (b) and (d) show the trend of VeCLIP with VeCap. The performance is improved significantly when we scale pre-training data up to 100M. Once over 100M, the performance gain becomes gradual and incremental.

Here we provide more examples of AltText and LLM-VeC from WIT-3M:

1. **AltText:** Ring Capri Pomellato | Pomellato Online Boutique
   VeCap: Pomellato's Ring Capri features a delicate and elegant white stone or possibly three pearls, set against a white background.
2. **AltText:** Fiamma F45 L 450 Royal Blue Awning.
   VeCap: The Fiamma F45 L 450 Royal Blue Awning is featured on a white car with a visible red logo for perfect closing, parked in a driveway under a tree, with a house in the background.
3. **AltText:** Union votes for strike on pensions
   VeCap: The man with white hair, dressed in a suit and tie, exhibits a surprised or expressive look on his face, with his mouth open and hand near his face, creating a dynamic and energetic expression.

Table A11: Data analysis on number of unique verbs, adjectives, nouns, and average token lengths with randomly sampled 10k data.

Caption	Verbs	Adjectives	Nouns	Avg. Length
AltText	1976	2098	13536	17.1
VeCap	2715	3697	20716	44.9
Gain	**+37.3%**	**+76.2%**	**+53.0%**	**+162.6%**

4. **AltText:** r/reallifedoodles - I can show you the world
   VeCap: The large orange and black drone hovers in the air, carrying two small teddy bears attached to it, above a patio area, as seen in the image.
5. **AltText:** 20 Amazon Skincare Products That Keep Selling Out
   VeCap: 20 Amazon skincare products that keep selling out feature a happy woman with dark skin, wearing a white shirt and covering her face with her hands, with a white spot or patch on her skin.
6. **AltText:** Durable White Arcane Dining Console Table With 6 Hidden Chairs
   VeCap: A durable white arcane dining console table with 6 hidden chairs is visually appealing and ready for use, as seen in the image featuring a dining set with a white table and two benches, surrounded by black chairs.
7. **AltText:** Peaceful apartment with wi fi internet access, near old Quebec.
   VeCap: Experience a peaceful stay in a cozy apartment with Wi-Fi internet access, located near historic Old Quebec, featuring a charming dining room with a set table and chairs on a hardwood floor, complete with a white refrigerator in the background.
8. **AltText:** CABLE BUJIA CHEVROLET CORSA 1.0 1.4 EFI FERRAZZI CABLE BUJIA CHEVROLET CORSA 1.0 1.4 EFI FERRAZZI
   VeCap An array of cords and wires, comprising a black rubber cable, is displayed on a pristine surface, featuring diverse configurations and orientations, with some lying horizontally and others positioned at angles.

Here we provide more examples of original caption and VeCap from CC3M:

1. **CC3M Caption:** person runs with the ball during their training session on friday.
   VeCap: A group of soccer players, clad in red and black jerseys, are energetically engaging in a game on a vast field, with some running and others immersed in the action, dispersed across the terrain.
2. **CC3M Caption:** a house with red roof with some bushes and a lamp post in front.
   VeCap: A prominent two-story beige building with a distinctive tile roof stands out in the area, illuminated by a nearby lamp post. The building appears to be a complex with several houses or apartments, adding a touch of complexity to the surroundings.

3. **CC3M Caption:** eating a big sweet cupcake with chocolate at cafe.
   VeCap: A person holds a half-eaten blueberry muffin on a plate, standing next to a dining table with a cup, while eating a big sweet cupcake with chocolate at a cafe.
4. **CC3M Caption:** paper heart with red ribbon and a bow.
   VeCap: A pink background showcases a heart-shaped box with a bow, adorned in white with the message "Happy Valentine's Day," positioned centrally within the image.
5. **CC3M Caption:** person andactor at the premiere
   VeCap: Two individuals, a man and a woman, are depicted standing together, both attired in formal attire. The man is donning a tuxedo with a black bow tie, while the woman is wearing a long dress. They seem to be positioning themselves for a photograph, possibly at a formal event.
6. **CC3M Caption:** wedding ceremony on the beach
   VeCap: A picturesque wedding ceremony unfolds on a stunning white sandy beach, where perfectly arranged chairs accommodate guests in formal attire. The groom and bride exude joy and love, basking in the warm sunlight.
7. **CC3M Caption:** revenge is a dish best served cold ... with lots of lettuce .
   VeCap: A large, possibly turtle, tortoise with an angry expression sits on rocks, displaying a saying or text message that reads "Revenge is a dish best cold served with lots of lettuce."
8. **CC3M Caption:** interior of an abandoned factory
   VeCap: The sunlit interior of an industrial building stands in contrast to its darker exterior, with numerous windows allowing natural light to flood the space, giving it an empty and open appearance devoid of people or personal touches.

Examining the aforementioned instances, it becomes evident that CC3M's captions exhibit a notable level of precision and high quality, displaying a closer alignment with the corresponding images. Conversely, WIT-3M's AltTexts tend to be more cluttered, signaling a comparatively subpar performance in contrast to CC3M. Upon implementing VeCap, even though CC3M's captions are of high quality, they are enhanced with more visual concepts leveraged via VeCap. Such integration of enriched visual concepts accounts for the significant improvement we achieve in retrieval tasks (the results are shown in Table A4).

F. More examples of WIT with VeCap

We conduct our scalable pipeline over 300 million image-text pairs. We randomly select more examples below to show the advantages of VeCap against the original AltText in terms of visual concepts. The examples are visualized in Figure A2.

Fig. A2: More examples of VeCap captions and AltTexts.