Contrast and Classify: Training Robust VQA Models (Supplementary)

Yash Kant[1*] Abhinav Moudgil[1] Dhruv Batra[1,2] Devi Parikh[1,2] Harsh Agrawal[1]
[1]Georgia Institute of Technology [2] Facebook AI Research

1. Ablations with Joint Training

In the joint training experiment (Table 2, Row 8), we use a weighing parameter (β) to combine the $\mathcal{L}_{\text{SC}}$ and $\mathcal{L}_{\text{CE}}$ losses. We ablate on the choice of weight (β) used, and we represent the overall loss in this experiment as:

$$\mathcal{L}_{joint} = \beta\mathcal{L}_{\text{SSC}} + (1-\beta)\mathcal{L}_{\text{CE}}$$

We also find that the VQA-Accuracy and Consensus Scores hit a sweet-spot at $\beta = 0.5$ and we use this configuration as our basline.

	Model	β	CS(4)	VQA val
1	MMT	0.25	52.97	66.14
2	MMT	0.50	53.63	66.23
3	MMT	0.75	48.53	61.34
4	MMT	0.90	40.68	51.03
5	MMT + ConClaT	-	**53.99**	**66.98**

Table A. Ablations on the choice of our hyper-parameter β for joint training.

2. Joint and Pretrain-Finetune Training

As mentioned in Section 4.3 of the manuscript, we respectively provide the training schemes used to jointly optimize in Algorithm 1 and the scheme used to pretrain-finetune in Algorithm 2 with the $\mathcal{L}_{\text{SSC}}$ and $\mathcal{L}_{\text{CE}}$ losses.

3. Gradient Surgery of $\mathcal{L}_{\text{SSC}}$ and $\mathcal{L}_{\text{CE}}$

To know whether the gradients of both the losses ($\mathcal{L}_{\text{SSC}}$ and $\mathcal{L}_{\text{CE}}$) are aligned with each other during training, we follow the gradient surgery setup of [8] for multi-task learning. During joint-training, we take the dot-products of gradients from both the losses and plot them to see how well they are aligned *i.e.* whether the dot product is positive or negative. In Figure A we plot the un-normalized dot product between the gradients corresponding to $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{SSC}}$ losses. We find that except for initial few steps the gradients of both the losses are aligned (dot product is positive) and thus the updates are complementary with respect to each other.

Algorithm 1 ConClaT with joint $\mathcal{L}_{\text{SSC}}$ and $\mathcal{L}_{\text{CE}}$

input: steps N; constant N_r, β; data $\mathcal{D}^{\text{aug}}$; networks f, g

for all $i \in \{1, \ldots, N\}$ **do**
  # initialize batches
  $\mathcal{B}_{\text{SSC}} = \text{CURATE}(N_r, \mathcal{D}^{\text{aug}}, \boldsymbol{w})$; $\mathcal{B}_{\text{CE}} \sim \mathcal{D}^{\text{aug}}$
  # compute gradients separately
  $\nabla_{\text{SSC}} = \nabla\, \mathcal{L}_{\text{SSC}}(f, g, \mathcal{B}_{\text{SSC}}) \cdot \beta$
  $\nabla_{\text{CE}} = \nabla\, \mathcal{L}_{\text{CE}}(f, g, \mathcal{B}_{\text{CE}}) \cdot (1-\beta)$
  # joint update
  update $f(.), g(.)$ networks with $\nabla = \nabla_{\text{CE}} + \nabla_{\text{SSC}}$
return network $f(.)$; throw away $g(.)$

Algorithm 2 ConClaT with pre-train $\mathcal{L}_{\text{SSC}}$ and fine-tune $\mathcal{L}_{\text{CE}}$

input: steps N_p, N_f; data $\mathcal{D}^{\text{aug}}$; networks f, g
  # pretrain with SSCL
  for all $i \in \{1, \ldots, N_p\}$ **do**
    $\mathcal{B} = \text{CURATE}(N_r, \mathcal{D}^{\text{aug}}, \boldsymbol{w})$
    update $f(.), g(.)$ networks to minimize $\mathcal{L}_{\text{SSC}}$ over $\mathcal{B}$
  # finetune with CE
  for all $i \in \{1, \ldots, N_f\}$ **do**
    $\mathcal{B} \sim \mathcal{D}^{\text{aug}}$
    update $f(.)$ network to minimize $\mathcal{L}_{\text{CE}}$ over $\mathcal{B}$
return network $f(.)$; throw away $g(.)$

4. Training

Hyperparameters. All the models have ∼100M trainable parameters. We train our models using Adam optimizer [3] with a linear warmup and with a learning rate of 1e-4 and a staircase learning rate schedule, where we multiply the learning rate by 0.2 at 10.6K and at 15K iterations. We train for 5 epochs of augmented train dataset $\mathcal{D}^{aug}$ on 4 NVIDIA Titan XP GPUs and use a batch-size of 420 when using $\mathcal{L}_{\text{SSC}}$ and $\mathcal{L}_{\text{CE}}$ both and 210 otherwise. We use PyTorch [4] for all the experiments. Hyperparameters

*Correspondence to ysh.kant@gmail.com

Table B. Hyperparameter choices for models.

#	Hyperparameters	Value	#	Hyperparameters	Value
1	Maximum question tokens	23	2	Maximum object tokens	101
3	$\mathcal{L}_{\text{CE}}$:$\mathcal{L}_{\text{SSC}}$ iterations ratio	3:1	4	Number of TextBert layers	3
5	Embedding size	768	6	Number of Multimodal layers	6
7	Multimodal layer intermediate size	3072	8	Number of attention heads	12
9	Negative type weights ($\boldsymbol{w}$)	$(0.25, 0.25, 0.5)$	10	Multimodal layer dropout	0.1
11	Similarity Threshold (ϵ)	0.95	12	Optimizer	Adam
13	Batch size	210/420	14	Base Learning rate	2e-4
15	Warm-up learning rate factor	0.1	16	Warm-up iterations	4266
17	Vocabulary size	3129	18	Gradient clipping (L-2 Norm)	0.25
19	Number of epochs	5/20	20	Learning rate decay	0.2
21	Learning rate decay steps	10665, 14931	22	Number of iterations	25000
23	Projection Dimension ($\mathcal{R}^{d_z}$)	128	24	Scaling Factor (s)	20
25	N_{ce}	4	26	N_r	70

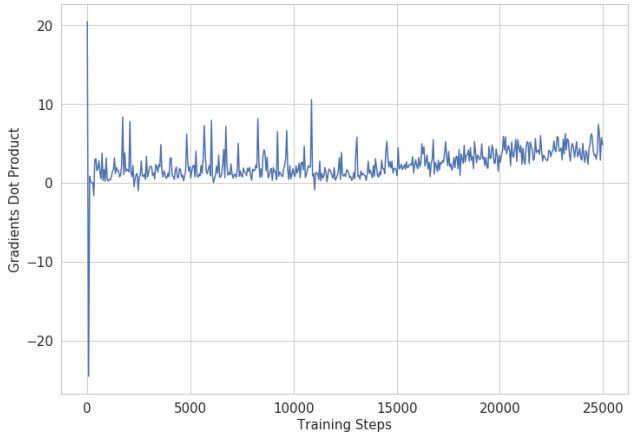


Figure A. Gradient Alignment between the $\mathcal{L}_{\text{SSC}}$ and $\mathcal{L}_{\text{CE}}$ losses. The dot-product is positive indicating that the gradients from the two losses are aligned.

are summarized in Table B.

5. Frequently Asked Questions

Why use different sampling rates for different negative types? The different types of negatives – same-image-different-question (`img`) and same-question-different-image (`que`) – encourages the model to be sensitive to both modalities. We use different sampling weights to emphasize more on these two types of negatives over the ones which just have different answers. We obtain the weights (Table B, Row 9) through hyper-parameter tuning on the validation set.

Why should questions dealing with different concepts but same answer (e.g., questions in Fig 2b, "Is the dog atop a sofa?" and "Is there broccoli in the picture?") have similar representations? We clarify that we do not impose any supervision at the level of MMT layers but only at the penultimate layer before answer prediction. Hence, the model is able to perform different reasoning steps (needed to process entirely different visual/textual inputs) for arriving at the same final answer.

Comparison of parameters, model size, training and inference time of ConClaT with baseline. Baseline (VQA model trained with augmented data) and ConClaT architectures are identical during inference and both have ∼135M parameters. An additional projection head (4M params) is used in ConClaT training for contrastive learning, which is discarded after training. Our model checkpoint consumes 1.5 GB of memory. Both ConClaT and baseline are trained for 25K iterations and takes roughly the same amount of time – 18 hours of wall-clock time on 4 TitanX GPUs. We do not find a significant difference in training times of ConClaT and the baseline. Evaluation times on the VQA test-set (443,757 samples) are also identical – nearly 10 minutes of wall-clock time for both ConClaT and baseline.

Joint training vs alternate training, which one is better? As evident from supplementary Table A, joint training is very sensitive to the weight hyperparameter (β). We hypothesize that this prevents joint-training from being as effective as ConClaT and so, we suggest alternate training as a better choice.

How was $N_{ce}(=4)$ chosen in ConClaT? We found similar performance with $N_{ce} = 2, 3, 4$. We use $N_{ce} = 4$ for faster training.

Does ConClaT improve using any dynamic word embeddings, such as ELMo/BERT? Similar to ELMo, our method already encodes questions using contextual word embeddings from a pretrained BERT model. Tempted by the question, we trained a variant of ConClaT with a randomly initialized BERT model. We find that gains from ConClaT are more significant (+2.53% vs +1.53% from the

paper) for **CS(4)** under this setting.

6. Augmented Data

Back-translation: We use 88 different MarianNMT [1] Back-translation model pairs released by Hugging Face [7] to generate question paraphrases. We use Sentence-BERT [5] to filter out paraphrases that cosine similarity of ≥ 0.95 with the original question and choose three unique paraphrases randomly from the filtered set. After filtering duplicates we end up with 2.89 paraphrases per original question on average.

VQG: We use the VQG model introduced by previous work [6] that takes as input the image and answer to generate a paraphrased question. We input the VQG module with 88 random noise vectors to keep the generation comparable with Back-translation approach. For filtering, we use the gating mechanism used by the authors and sentence similarity score of ≥ 0.85 and keep a maximum of 3 unique rephrasings for each question. Since, VQG produces fewer unique rephrasings per question than Back-translation, we used a lower similarity threshold. After filtering duplicates we end up with only 0.96 paraphrases per original question on average, far fewer than Back-translation. Qualitatively, we find the VQG paraphrases worse when compared against Back-translated ones.

Evaluation: During training, we evaluate our models using the Back-translated rephrasings on a subset of questions from validation set which do not overlap with VQA-Rephrasings [6] dataset.

7. Code and Result Files

We share the code for running the baseline and the best experiments (Table 1, Rows 5, 9). Please find the released code at: https://www.github.com/yashkant/concat-vqa

8. Full Ablations

For brevity and conciseness, we omitted **CS(1)** and **CS(2)** scores in the main ablation table, we provide the these scores in Table B.

9. Qualitative Samples

Figures B, C, D, E show many more qualitative samples comparing the baseline and ConClaT. We visualize the data generated via Back-translation and mined triplets in Figures F, G, H.

References

[1] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics. 3

[2] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, A. Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *ArXiv*, abs/2004.11362, 2020. 4

[3] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. 1

[4] Adam Paszke, S. Gross, Francisco Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, Alban Desmaison, Andreas Köpf, E. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, B. Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *ArXiv*, abs/1912.01703, 2019. 1

[5] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. 3

[6] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6649–6658, 2019. 3

[7] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019. 3

[8] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning, 2020. 1

[9] Yiyi Zhou, Rongrong Ji, Jinsong Su, Xiangming Li, and Xiaoshuai Sun. Free vqa models from knowledge inertia by pairwise inconformity learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9316–9323, Jul. 2019. 4

	Model	Loss(es)	Scaling	N-Type	Train Scheme	CS(1)	CS(2)	CS(3)	CS(4)	VQA val
1	MMT	$\mathcal{L}_{\text{CE}}$	-	-	-	67.58	60.04	55.53	52.36	66.31
2	MMT	$\mathcal{L}_{\text{SSC}}$ & $\mathcal{L}_{\text{CE}}$	✓	R	Alternate	68.19	60.92	56.53	53.42	66.62
3	MMT	$\mathcal{L}_{\text{SC}}$ & $\mathcal{L}_{\text{CE}}$	✓	RQ	Alternate	68.41	61.24	56.88	53.77	66.97
4	MMT	$\mathcal{L}_{\text{SSC}}$ & $\mathcal{L}_{\text{CE}}$	✓	RI	Alternate	68.47	61.28	56.91	53.79	66.93
5	MMT	$\mathcal{L}_{\text{SSC}}$ & $\mathcal{L}_{\text{CE}}$	✓	QI	Alternate	68.65	61.40	57.00	53.90	66.95
6	MMT	$\mathcal{L}_{\text{SSC}}$ & $\mathcal{L}_{\text{CE}}$	✓	RQI	Alternate	**68.62**	**61.42**	**57.08**	**53.99**	**66.98**
7	MMT	$\mathcal{L}_{\text{SC}}$ & $\mathcal{L}_{\text{CE}}$	✗	RQI	Alternate	68.20	60.90	56.49	53.36	66.60
8	MMT	$\mathcal{L}_{\text{SSC}}$ & $\mathcal{L}_{\text{CE}}$	Dynamic	RQI	Alternate	68.60	61.38	57.01	53.92	66.95
9	MMT	$\mathcal{L}_{\text{SSC}}$ & $\mathcal{L}_{\text{CE}}$	✓	RQI	Joint	67.75	60.79	56.59	53.63	66.23
10	MMT	$\mathcal{L}_{\text{SSC}} \rightarrow \mathcal{L}_{\text{CE}}$ [2]	✗	RQI	Pretrain-Finetune	65.33	57.39	52.63	49.20	64.21
11	MMT	$\mathcal{L}_{\text{DMT}}$ [9] & $\mathcal{L}_{\text{CE}}$	✗	RQI	Alternate	68.11	60.70	56.23	53.10	66.59

Table B. **Ablations Study of ConClaT**. **Scaling** denotes whether scaling factor α was used. **N-Type** defines the type of negatives used from Image (I), Question (Q) and Random (R). All experiments are run with Back Translation data.

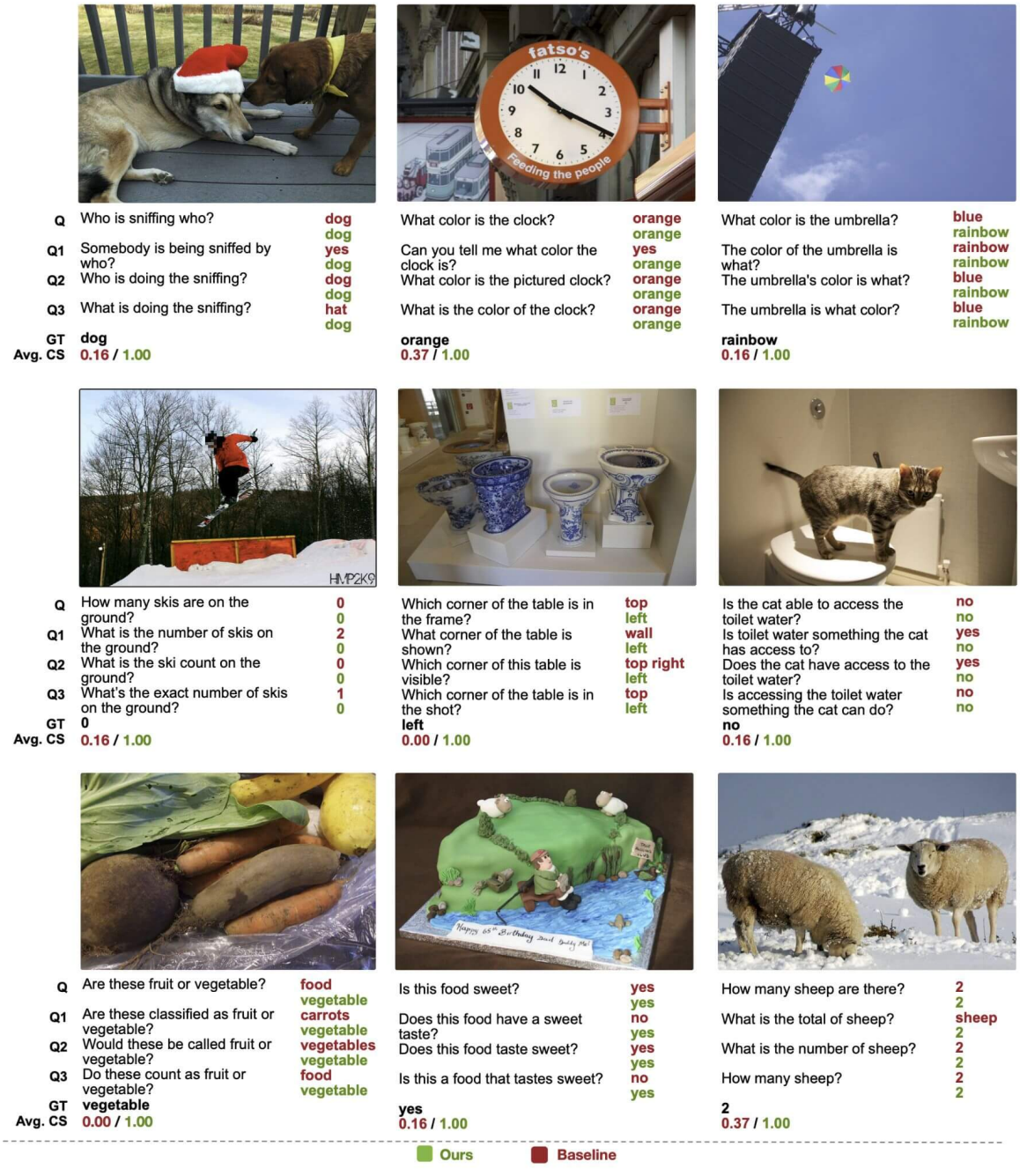


Figure B. Qualitative Examples. Predictions of ConClaT and MMT+CE baseline on several image-question pairs and their corresponding rephrased questions. Average Consensus Scores (k=1-4) are also shown at the bottom (higher the better).

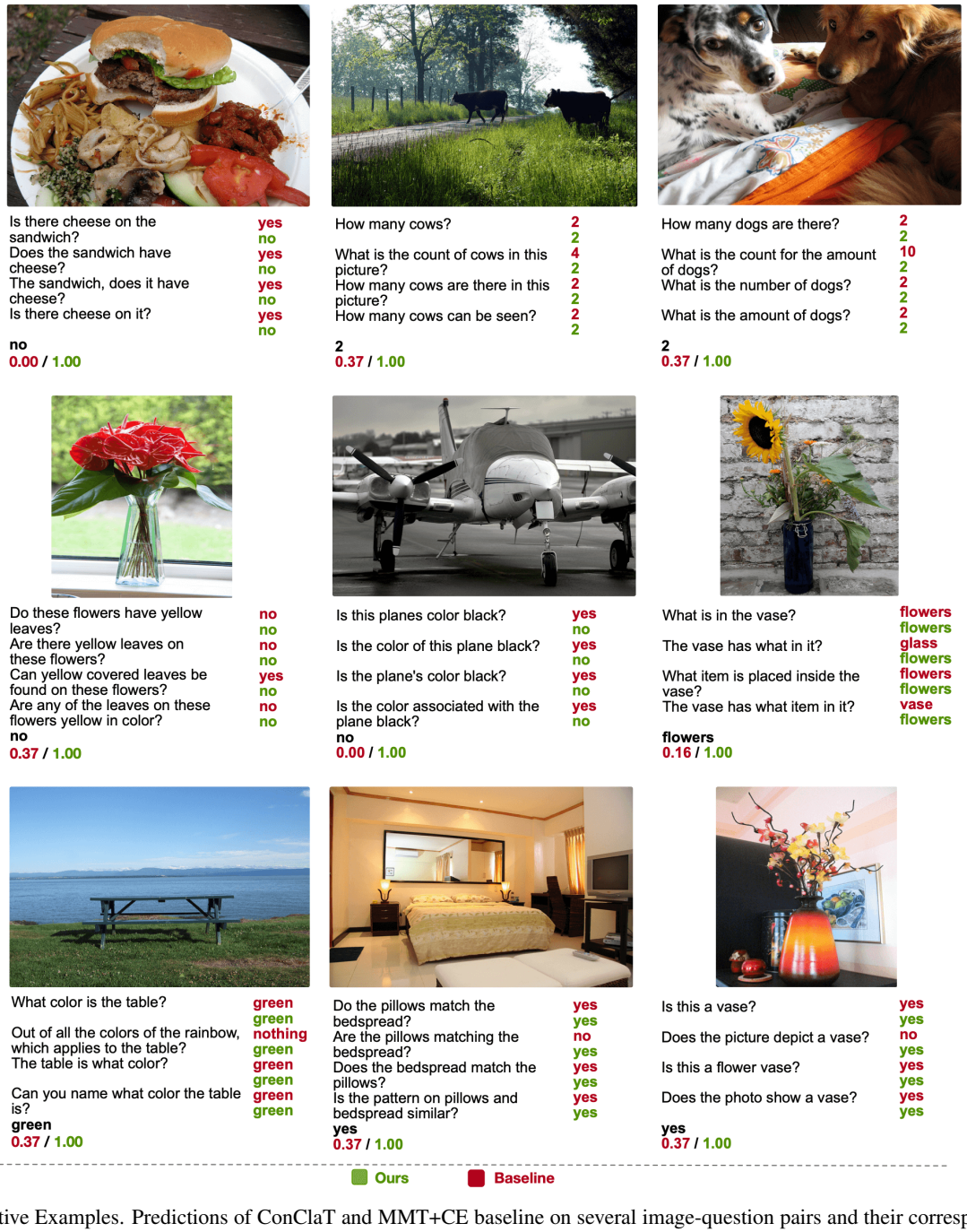


Figure C. Qualitative Examples. Predictions of ConClaT and MMT+CE baseline on several image-question pairs and their corresponding rephrased questions. Average Consensus Scores (k=1-4) are also shown at the bottom (higher the better).

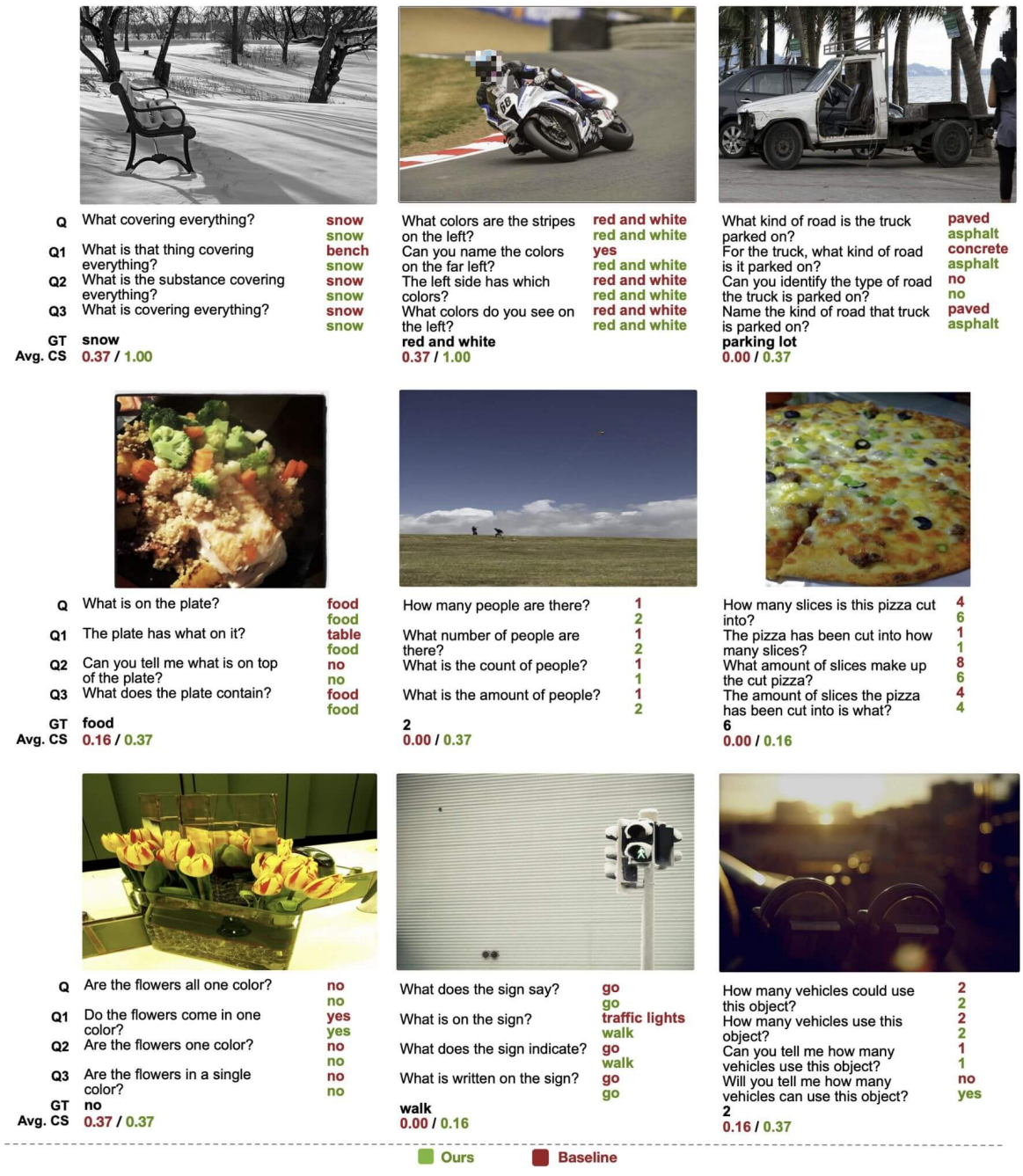


Figure D. Qualitative Examples. Predictions of ConClaT and MMT+CE baseline on several image-question pairs and their corresponding rephrased questions. Average Consensus Scores (k=1-4) are also shown at the bottom (higher the better).

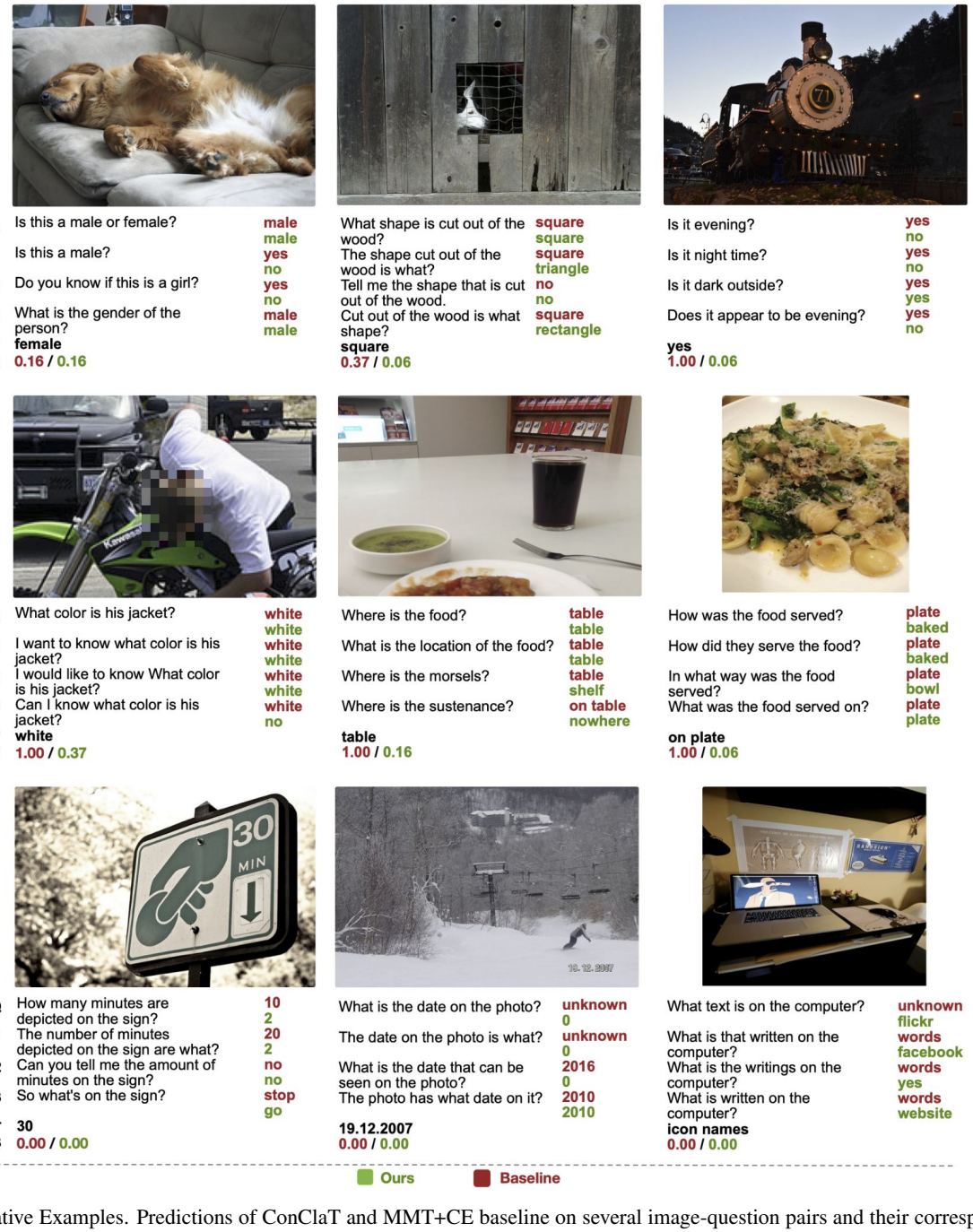


Figure E. Qualitative Examples. Predictions of ConClaT and MMT+CE baseline on several image-question pairs and their corresponding rephrased questions. Average Consensus Scores (k=1-4) are also shown at the bottom (higher the better).

	Reference Sample	Image Negative	Question Negative
Image			
Original Question	Where is the dog laying?	What is the dog doing?	The dog's lying on a rug?
Back Translated Questions	Where's the dog lay? Where's the dog lying down? Where's the dog lying?	What's the dog doing? What is a dog doing? What's that dog doing?	Is the dog laying on a rug? Is the dog lying on a rug? Is the dog lying on a carpet?
Ground Truth Answers	outside, street, yes, sidewalk, ground	resting, lying down, laying down, sleeping	yes
Image			
Original Question	What are the men sitting on?	What are the ranks of the military members?	What are men sitting on?
Back Translated Questions	What are these men sitting on? What are men sitting on? What were these men sitting on?	What are the ranks of military members? What are the ranks of the military? What is the rank of military members?	What are the men sitting on? What are these men sitting on? What were these men sitting on?
Ground Truth Answers	bench	low	bed
Image			
Original Question	What's in the oven?	What storage is open?	What is in the oven?
Back Translated Questions	What is in the oven? What is there in the oven? What is inside the oven?	What kind of storage is open? Which storage is open? What storage place is open?	What's in the oven? What was in the oven? What is the meaning of the oven?
Ground Truth Answers	pot	oven	turkey

Figure F. Visualizing the triplets of samples from VQA dataset with corresponding mined Image and Question Negatives.

	Reference Sample	Image Negative	Question Negative
Image			
Original Question	How high is the plane in the sky?	Are there clouds?	What's the altitude of the plane?
Back Translated Questions	How high is the plane in the air? How high is the plane of the sky? How high is the flying plane?	Are There Clouds? Is there clouds? Is there a clouds?	What is the altitude of the plane? What is the plane's altitude? How high is the altitude of the plane?
Ground Truth Answers	medium, very, high, very high	yes	unknown, high
Image			
Original Question	What is the woman taking a picture of?	What color is the grass?	Is that woman taking pictures?
Back Translated Questions	What's the woman taking a picture of? What is the woman taking a picture of it? What's that woman taking a picture of?	What color is grass? What color is this grass? What color does the grass have?	Is this woman taking a picture? Is the woman taking a picture? Does this woman take a picture?
Ground Truth Answers	goose, swan, bird	green, no grass	yes
Image			
Original Question	Is the catcher wearing safety gear?	What is the name of the teams?	Is the athlete wearing safety gear?
Back Translated Questions	Is the catcher wearing safety equipment? Is the catcher wearing the safety equipment? Does the catcher wear safety equipment?	What's the name of the teams? What's the name of these teams? What is the name of teams?	Is the athlete wearing a safety gear? Is the athlete wearing safety equipment? Does the athlete wear safety equipment?
Ground Truth Answers	yes	cubs	no

Figure G. Visualizing the triplets of samples from VQA dataset with corresponding mined Image and Question Negatives.

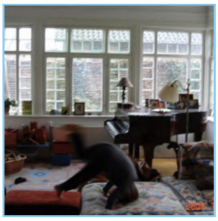
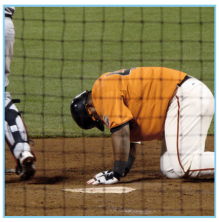
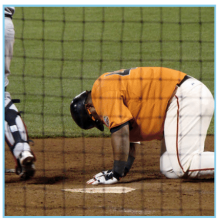
	Reference Sample	Image Negative	Question Negative
Image			
Original Question	What is the elephant doing?	How many elephants are there?	What is the elephant doing?
Back Translated Questions	What's the elephant doing? What's an elephant doing? What's that elephant doing?	How many elephants is there? How many elephants do you have? How many elephants exist?	What's the elephant doing? What's an elephant doing? What's that elephant doing?
Ground Truth Answers	drinking water, drinking	3, 4	eating, kissing, playing
Image			
Original Question	What is the equipment in the background?	How many birds?	What instrument is in the background?
Back Translated Questions	What's the equipment in the background? What are the equipment in the background? What kind of equipment is in the background?	How many birds are? How many birds are there? How much birds?	What instrument in the background? Which instrument is in the background? What is the instrument in the background?
Ground Truth Answers	building, oil, cranes, boats	3, 2	piano
Image			
Original Question	What brand are the catcher's shoes?	What game are they playing?	What brand are the batter's shoes?
Back Translated Questions	What brand is the catcher's shoes? What brand are the shoes of the catcher? What brand are the catcher shoes?	What kind of game are they playing? What game do they play? Which game are they playing?	What are the batter's shoes brand? What brand are the shoes of the batter? What brand are the batter shoes?
Ground Truth Answers	puma, black and white, asics, nike	baseball	clear, adidas, nike

Figure H. Visualizing the triplets of samples from VQA dataset with corresponding mined Image and Question Negatives.