Large-scale Pretraining for Visual Dialog: A Simple State-of-the-Art Baseline

Vishvak Murahari¹ Dhruv Batra^{1,2} Devi Parikh^{1,2} Abhishek Das^{$1\rightarrow 2$}

¹ Georgia Institute of Technology ² Facebook AI Research {vishvak.murahari,dbatra,parikh,abhshkdz}@gatech.edu

Abstract. Prior work in visual dialog has focused on training deep neural models on VisDial [1] in isolation. Instead, we present an approach to leverage pretraining on related vision-language datasets before transferring to visual dialog. We adapt the recently proposed ViLBERT model [2] for multi-turn visually-grounded conversations. Our model is pretrained on the Conceptual Captions [3] and Visual Question Answering [4] datasets, and finetuned on VisDial. Our best single model outperforms prior published work by > 1% absolute on NDCG and MRR.

Next, we find that additional finetuning using "dense" annotations in VisDial leads to even higher NDCG – more than 10% over our base model – but hurts MRR – more than 17% below our base model! This highlights a trade-off between the two primary metrics – NDCG and MRR – which we find is due to dense annotations not correlating well with the original ground-truth answers to questions.

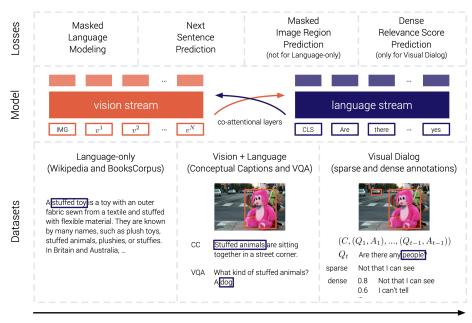
Keywords: Vision & Language, Visual Dialog

1 Introduction

Recent years have seen incredible progress in Visual Dialog [1,5–25], spurred in part by the initial efforts of Das *et al.* [1] in developing a concrete task definition – given an image, dialog history consisting of a sequence of question-answer pairs, and a follow-up question about the image, to predict a free-form natural language answer to the question – along with a large-scale dataset and evaluation metrics. The state-of-the-art on the task has improved by more than 20% absolute ($\sim 54\% \rightarrow \sim 74\%$ NDCG) and the original task has since been extended to challenging domains, *e.g.* video understanding [26], navigation assistants [27–29].

While this is promising, much of this progress has happened in isolation, wherein sophisticated neural architectures are trained and benchmarked solely on the Vis-Dial dataset. This is limiting – since there is a significant amount of shared abstraction and visual grounding in related tasks in vision and language (e.g. captioning, visual question answering) that can benefit Visual Dialog – and wasteful – since it is expensive and dissatisfying to have to collect a large-scale dataset for

Vishvak Murahari et al.



Training Curriculum

Fig. 1: First, the language stream of our model is pretrained on English Wikipedia and the BooksCorpus [30] datasets with the masked language modeling (MLM) and next sentence prediction (NSP) losses. Next, the entire model is trained on the Conceptual Captions [3] and VQA [4] datasets with the masked image region (MIR), MLM and NSP losses. Finally, we finetune the model on sparse annotations from VisDial [1] with the MIR, MLM and NSP losses, and optionally finetune on dense annotations.

every new task. In this work, we explore an approach to pretrain our model on other related vision and language datasets and then transfer to Visual Dialog.

Our work is inspired by prior work in transfer learning in computer vision and natural language understanding where large models [31–41] are pretrained on large datasets [30, 42, 43] with simple yet powerful self-supervised objectives to learn powerful representations that are then transferred to downstream tasks, leading to state-of-the-art results on a variety of benchmarks [42,44]. Recent work has extended this to vision and language tasks [2, 45–50], leading to compelling results in Visual Question Answering [4], Commonsense Reasoning [51], Natural Language Visual Reasoning [52], Entailment [53], Image-Text Retrieval [54, 55], Referring Expressions [56], and Vision-Language Navigation [57].

In this work, we adapt ViLBERT [2] to Visual Dialog. ViLBERT uses two Transformer-based [35] encoders, one for each of the two modalities – language and vision – and interaction between the two modalities is enabled by coattention layers *i.e.* attention over inputs from one modality conditioned on inputs from the other. Note that adapting ViLBERT to Visual Dialog is not trivial. The Visual Dialog dataset has image-grounded conversation sequences that

are up to 10 rounds long. These are significantly longer than captions (which are ≤ 2 sentences) from the Conceptual Captions dataset [3] or question-answer pairs from VQA [4] used to pretrain ViLBERT, and thus requires a different input representation and careful reconsideration of the masked language modeling and next sentence prediction objectives used to train BERT [36] and ViLBERT [2].

This adapted model outperforms prior published work by > 1% absolute and achieves state-of-the-art on Visual Dialog. Next, we carefully analyse our model and find that additional finetuning on 'dense' annotations¹ *i.e.* relevance scores for all 100 answer options corresponding to each question on a subset of the training set, highlights an interesting trade-off – the model gets to ~74.5% NDCG (outperforming the 2019 VisDial Challenge winner), but an MRR of ~52% (~17% below our base model!). We find this happens because dense annotations in VisDial do not correlate well with the ground-truth answers to questions, often rewarding the model for generic, uncertain responses.

Concretely, our contributions are as follows:

- We introduce an adaptation of the ViLBERT [2] model for Visual Dialog, thus making use of the large-scale Conceptual Captions [3] and Visual Question Answering (VQA) [4] datasets for pretraining and learning powerful visually-grounded representations before finetuning on VisDial [1]. Since captioning and VQA differ significantly from Visual Dialog in input size (≤ 2 sentence descriptions $vs. \leq 10$ question-answer rounds), this requires rethinking the input representation to learn additional segment embeddings representing questions-answer pairs. Our adapted model improves over prior published work by > 1% and sets a new state-of-the-art.
- We next finetune our model on dense annotations *i.e.* relevance scores for all 100 answer options corresponding to each question on a subset of the training set, leading to even higher NDCG more than 10% over our base model but hurting MRR more than 17% below our base model! This highlights a stark trade-off between the two primary metrics for this task NDCG and MRR. Through qualitative and quantitative results, we show that this happens because dense annotations do not correlate well with the original ground-truth answers, often rewarding the model for generic, uncertain responses.
- Our PyTorch [58] code is publicly available² to encourage further work in large-scale transfer learning for VisDial.

2 Related Work

Our work is related to prior work in visual dialog [1, 5-25], and self-supervised pretraining and transfer learning in computer vision and language [31-41].

¹ publicly available on visualdialog.org/data.

² github.com/vmurahari3/visdial-bert/

Visual Dialog. Das *et al.* [1] and de Vries *et al.* [5] introduced the task of Visual Dialog – given an image, dialog history consisting of a sequence of questionanswer pairs, and a follow-up question, predict a free-form natural language answer to the question – along with a dataset, evaluation metrics, and baseline models. Follow-up works on visual dialog have explored the use of deep reinforcement learning [6, 7, 20], knowledge transfer from discriminative to generative decoders [8], conditional variational autoencoders [9], generative adversarial networks [10], attention mechanisms for visual coreference resolution [12, 14], and modeling the questioner's theory of mind [13]. Crucially, all of these works train and evaluate on the VisDial dataset *in isolation*, without leveraging related visual grounding signals from other large-scale datasets in vision and language. We devise a unified model that can be pretrained on the Conceptual Captions [3] and VQA [4] datasets, and then transferred and finetuned on VisDial.

Self-Supervised Learning in Vision and Language. Building on the success of transfer learning in natural language understanding [34–41] leading to stateof-the-art results on a broad set of benchmarks [42,44], recent work has extended this to vision and language tasks [2,45–50]. These works pretrain single [45,48,49] or two [2,46]-stream Transformer [35]-based models with self-supervised objectives, such as next-sentence prediction and masked language/image modeling, on large-scale image-text datasets and have led to compelling results in Visual Question Answering [4], Commonsense Reasoning [51], Natural Language Visual Reasoning [52], Entailment [53], Image-Text Retrieval [54,55], and Referring Expressions [56], and Vision-Language Navigation [57].

3 Adapting ViLBERT [2] for Visual Dialog

Lu *et al.* [2] introduced ViLBERT³, which extended BERT [36] to a two-stream multi-modal architecture for jointly modeling visual and linguistic inputs. Interaction between the two modalities was enabled through co-attention layers, *i.e.* attending to one modality conditioned on the other – attention over language conditioned on visual input, and attention over image regions conditioned on linguistic input. This was operationalized as swapping the key and value matrices between the visual and linguistic Transformer [35] blocks. We next discuss our changes to adapt it for Visual Dialog followed by our training pipeline.

Input Representation. Recall that the model gets image I, dialog history (including image caption C) $H = (C, (Q_1, A_1), ..., (Q_{t-1}, A_{t-1}))$, question Q_t , and a list of 100 answer options $A_t = \{A_t^{(1)}, A_t^{(2)}, ..., A_t^{(100)}\}$ as input, and is asked to return a sorting of A_t . We concatenate the t rounds of dialog history and follow-up question Q_t , with each question and answer separated by a <SEP> token. The overall input to the language stream is represented as:

 $\langle \text{CLS} \rangle C \langle \text{SEP} \rangle Q_1 \langle \text{SEP} \rangle A_1 \langle \text{SEP} \rangle, \dots, \langle \text{SEP} \rangle Q_t \langle \text{SEP} \rangle A_t \langle \text{SEP} \rangle$ (1)

³ along with code released at github.com/jiasenlu/ViLBERT_beta.

Similar to Wolf *et al.* [59], we use different segment embeddings for questions and answers to help the model distinguish between the two and understand question and answer boundaries in the input. Captions and answers share the same segment embeddings. To represent the image, we follow [2, 60] and extract object bounding boxes and their visual features for top-36 detected objects in the image from a Faster R-CNN [61] (with a ResNet-101 [31] backbone) object detection network pretrained on the Visual Genome dataset [43]. The feature vector for each detected object is computed as mean-pooled convolutional features from the regions of that object. A 5-d feature vector, consisting of normalized top-left and bottom-right object coordinates, and the fraction of image area covered, is projected to the same dimensions as the feature vector for the detected object, and added to it, giving us the final visual features { $v_1, ..., v_{36}$ }. The beginning of this image region sequence (consisting of object detection features) is demarcated by an IMG token with mean-pooled features from the entire image. The overall input to ViLBERT can be written as the following sequence:

 $v_1, ..., v_{36}$ <CLS> C <SEP> Q_1 <SEP> A_1 <SEP> , ..., <SEP> Q_t <SEP> A_t <SEP> (2)

3.1 Pretraining on Conceptual Captions [3]

To pretrain the model, we follow [2] and train on the Conceptual Captions (CC) dataset, which is a large corpus (with $\sim 3M$ samples) of aligned image-caption pairs. During pretraining, the sum of the masked language modeling (MLM) loss [36] and the masked image region (MIR) loss is optimized. To compute the MLM loss, a set of tokens in the input sequence are masked and the model is trained to predict these tokens given context. We mask around 15% of the tokens in the input sequence. For the MIR loss, similar to the MLM loss, we zero out 15% of the image features and the model learns to predict the semantic category of the masked out object (out of 1601 classes from Visual Genome [43, 60]).

3.2 Pretraining on VQA [4]

The VQA dataset is quite related to Visual Dialog in that it can be interpreted as independent visually-grounded question-answer pairs with no dialog history, and thus is a natural choice for further pretraining prior to finetuning on VisDial. Similar to Lu *et al.* [2], we pretrain on VQA by learning a small decoder – a two-layer MLP – on top of the element-wise product between the image and text representations to predict a distribution over 3129 answers.

3.3 Finetuning on Visual Dialog [1]

To finetune on Visual Dialog, we use the MLM loss along with the next sentence prediction (NSP) and MIR losses. For MLM, we mask 10% of the tokens in the dialog sequence. For MIR, similar to pretraining, we mask 15% of the image features. Note that the discriminative task in visual dialog is to identify the ground-truth answer from a list of 100 answer options consisting of popular, nearest neighbors, and random answers from the dataset. We achieve this through the NSP loss. The NSP head is trained to predict 1 when the groundtruth answer is appended to the input sequence, and 0 when a negative answer sampled from the remaining answer options is appended to it. Each image in VisDial has 10 rounds of dialog, leading to 10 sets of positive and negative samples for the NSP loss per mini-batch. Since these are fairly correlated samples, we randomly sub-sample 2 out of these 20 during training. At test time, we use log-probabilities from the NSP head to rank the 100 answer options per round.

3.4 Finetuning with Dense Annotations

The authors of [1] recently released dense annotations⁴ *i.e.* relevance scores for all 100 answer options from A_t corresponding to the question on a subset of the training set. These relevance scores range from 0 to 1 and are calculated as the ratio of number of human annotators who marked a particular answer option as correct to the total number of human annotators (= 4). So 1 means that the answer option was considered correct by 4 human annotators. In our final stage of training, we utilize these dense annotations to finetune our model. Concretely, we use the NSP head to predict likelihood scores $\hat{\ell}_t^{(i)}$ for each answer option $A_t^{(i)}$ at round t, normalize these to form a probability distribution over the 100 answers $\hat{y}_t = [\hat{y}_t^{(1)}, ..., \hat{y}_t^{(100)}]$, and then compute a cross-entropy (CE) loss against the normalized ground-truth relevance scores y_t , given by $-\sum_i y_t^{(i)} \log \hat{y}_t^{(i)}$.

4 Experiments

To compare to previous research, we conduct experiments on VisDial v1.0 [1]. The dataset contains human-human dialogs on $\sim 130k$ COCO [62]-like images. We follow the original splits and use $\sim 120k$ for training, $\sim 2k$ for validation, and $\sim 8k$ for testing. We next describe the various settings we experiment with.

Evaluation Metrics. We use metrics introduced in [1]. Specifically, given the predicted ranking of 100 answer options from a model at each round, we compute retrieval metrics – mean rank (MR) of the ground-truth answer, mean reciprocal rank (MRR), and recall@k ($k = \{1, 5, 10\}$). Additionally, along with the release of dense annotations, *i.e.* relevance scores $\in [0, 1]$ for all 100 answer options, a new metric – NDCG – was introduced. NDCG accounts for multiple correct answers in the option set and penalizes low-ranked but correct answer options.

4.1 Language-only

We begin with a 'blind' setting, where given the dialog history and follow-up question, and without access to the image, the model is tasked with predicting

⁴ publicly available on visualdialog.org/data.

the answer. We do not use the ViLBERT formulation for these experiments, and finetune the BERT model released in [36] and pretrained on BooksCorpus [30] and English Wikipedia. For the MLM loss, we mask 15% of tokens and subsample 8 out of 20 sequences per mini-batch during training. We experiment with two variants – training only with NSP, and training with both NSP and MLM. See Table 3 for language-only results (marked 'L-only'). This setting helps us benchmark gains coming from switching to Transformer [35]-based architectures before the added complexity of incorporating visual input.

Varying number of dialog rounds. We train ablations of our language-only model (with NSP and MLM losses) where we vary the number of rounds in dialog history, starting from 0, where the input sequence only contains the follow-up question and answer, to 2, 4, and 6 and 10 rounds of dialog history (Table 1).

Zero-shot and 'cheap' finetuning. We report performance for ablations of our NSP+MLM model with no/minimal training in Table 2. First, we do a zero-shot test where we initialize BERT with weights from Wikipedia and BooksCorpus pretraining and simply run inference on VisDial. Second, with the same initialization, we freeze all layers and finetune only the MLM and NSP loss heads.

4.2 Finetuning on VisDial

We finetune ViLBERT on VisDial with four different weight initializations – 1) with randomly initialized weights, 2) from the best language-only weights (from Section 4.1) for the language stream (visual stream and co-attention layers initialized randomly), 3) from a model pretrained on CC [3] (as described in Section 3.1) and 4) from a model pretrained on CC [3]+VQA [4] (as described in Section 3.2). 1) helps us benchmark improvements due to pretraining, 2) helps us benchmark performance if the model learns visual grounding solely from Vis-Dial, 3) quantifies effects of learning visual grounding additionally from CC, and 4) helps us quantify improvements with additional exposure to visually-grounded question-answering data. See Table 3 for results.

4.3 Finetuning with Dense Annotations

Finally, we finetune our best model from Section 4.2 - marked 'w/ CC+VQA'in Table 3 – on dense annotations, as described in Section 3.4. Note that computing the CE loss requires a separate forward pass for each of the 100 answer options, since dialog history, question, answer are all concatenated together before passing as input. This is memory-expensive, and so in practice, we subsample and only use 80 options, and use gradient accumulation to (artificially) construct a larger mini-batch. Finetuning with the CE loss only leads to significant improvements on NDCG but hurts other metrics (see Table 3). We discuss and analyse this in more detail later. But to control for this 'metric-overfitting', we also train a variant with both the CE and NSP losses.

Vishvak Murahari et al.

5 Results

We list findings from all experiments described in Section 4 below.

Table 1: Performance of the NSP + MLM language-only model on VisDial v1.0 val as the number of dialog history rounds is varied

0	50.54	54.29	38.88	72.67	83.09	5.90
2	53.69	61.31	46.83	78.96	88.15	4.51
4	55.10	62.83	48.36	80.61	89.57	4.19
6	55.69	63.73	49.31	81.13	90.06	4.04
10	57.22	64.10	50.05	81.09	90.00	4.16

Table 2: Performance of the NSP + MLM language-only model on VisDial v1.0 val with no / minimal training (described in Sec. 4.1)

Model	NDCG ↑ I	MRR † I	R@1 ↑ 1	$R@5 \uparrow 1$	$R@10 \uparrow MR \downarrow$
No training	11.63	6.88	2.63	7.17	11.30 46.90
Loss heads only	19.69	9.81	3.42	10.44	$18.85 \ 31.38$

- Language-only performs well. The language-only model gets to 57.22 on NDCG and 64.10 on MRR (Table 3), which is already competitive with several prior published works (Table 4). These trends are consistent with high human performance on VisDial [1] with just language (question and dialog history) – 48.5 on MRR – which further improves to 63.5 on MRR with image.
- Increasing dialog history rounds helps. We report performance of the language-only model as a function of dialog history rounds in Table 1 and Fig. 2a. Note that the change in performance from including 0 to 4 rounds of dialog history (+4.56 on NDCG, +8.54 on MRR) is much more than from 4 to 10 dialog history rounds (+2.12 on NDCG, +1.27 on MRR). Thus, performance continues to go up with increasing dialog history rounds but starts to plateau with ≥ 4 history rounds. We believe these improvements are largely indicative of the Transformer's ability to model long-term dependencies.
- Zero-shot model performs poorly. Running inference with the languageonly model pretrained on BooksCorpus [30] and Wikipedia without any finetuning on VisDial only gets to 11.63 on NDCG and 6.88 on MRR (Table 2). Finetuning the loss heads with all other layers frozen leads to an improvement of ~8 NDCG points over this. This low performance can be attributed to significantly longer sequences in VisDial than the model was pretrained with.
- VQA initialization helps more than random or CC initialization. Finetuning ViLBERT on VisDial with weights initialized from VQA pretraining gets to 64.82 on NDCG and 68.97 on MRR, ~3 points better than random initialization on NDCG and ~4 points better than CC pretraining (Table 3). We

believe poorer transfer from CC is because both VQA and VisDial have images from COCO and are more closely related tasks than captioning on CC.

- Dense annotations boost NDCG, hurt MRR. Finetuning with the CE loss leads to 74.47 on NDCG a ~10% improvement over the 'w/ CC + VQA' base model but 50.74 on MRR, a ~17% decline below the base model (Table 4). This is a surprising finding! We carefully analyze this behavior in Section 6.
- Ensembling does not improve performance. We trained 3 models initialized with different random seeds for each of the 3 variants ('w/ CC + VQA', 'CE' and 'CE + NSP') and aggregated results by averaging the normalized scores from the 3 models. We did not observe any significant improvement.

Table 3: Results on VisDial v1.0 val (with 95% CI). ↑ indicates higher is better.

	Model	NDCG \uparrow	$\mathrm{MRR}\uparrow$	R@1 \uparrow	$R@5\uparrow$	R@10 \uparrow	$\mathrm{MR}\downarrow$
) Ily	NSP NSP + MLM	55.80 ±0.9	$63.37 {\ \pm 0.5}$	$49.28{\scriptstyle~\pm0.7}$	$80.51 {\ \pm 0.5}$	$89.22{\scriptstyle~\pm0.4}$	$4.32{\scriptstyle~\pm0.1}$
ر ۲	NSP + MLM	$57.22{\scriptstyle~\pm 0.9}$	$64.10{\scriptstyle \pm 0.5}$	$50.05{\scriptstyle~\pm0.7}$	$81.09 \ {\scriptstyle \pm 0.5}$	$90.00{\scriptstyle~\pm 0.4}$	$4.16{\scriptstyle~\pm 0.1}$
ਤ (Random init	$61.88{\scriptstyle~\pm 0.9}$	$67.04{\scriptstyle~\pm0.5}$	53.51 ± 0.7	$83.94 {\ \pm 0.5}$	92.27 ± 0.4	$3.55{\scriptstyle~\pm0.1}$
, isi	Random init w/ L-only w/ CC [3] w/ CC [3]+VQA [4]	61.21 ± 0.9	$66.75 \scriptstyle \pm 0.5$	$53.34{\scriptstyle~\pm0.7}$	$83.88{\scriptstyle~\pm0.5}$	$91.89{\scriptstyle~\pm 0.4}$	$3.68{\scriptstyle~\pm 0.1}$
- <u>-</u>	w/ CC [3]	60.80 ± 0.9	$67.21{\scriptstyle \pm 0.5}$	$53.36{\scriptstyle~\pm0.7}$	$84.39 \ {\scriptstyle \pm 0.5}$	$92.49 {}_{\pm 0.4}$	$3.44{\scriptstyle~\pm0.1}$
' (w/ CC [3]+VQA [4]	$64.82{\scriptstyle~\pm 0.9}$	68.97 ± 0.5	55.78 ±0.7	85.34 ± 0.5	$93.11 \scriptstyle \pm 0.4$	$3.35 {\scriptstyle \pm 0.1}$
) se	CE	75.10 ± 1.1	$52.12{\scriptstyle~\pm 0.6}$	$39.84{\scriptstyle~\pm0.7}$	$64.93 \scriptstyle \pm 0.7$	80.47 ± 0.5	$6.26{\scriptstyle~\pm0.1}$
$\left\{ \begin{array}{c} + det \\ + det \end{array} \right\}$	$\begin{array}{c} { m CE} { m CE} + { m NSP} \end{array}$	69.11 ±1.0	$65.76{\scriptstyle~\pm0.5}$	$53.30{\scriptstyle~\pm0.7}$	80.77 ± 0.5	$90.00{\scriptstyle~\pm0.4}$	$4.33{\scriptstyle~\pm0.1}$

We report results from the Visual Dialog evaluation server⁵ for our best models – 'w/ CC + VQA', 'CE' and 'CE + NSP' – on the unseen test-std split in Table 4. We compare against prior published results and top entries from the leaderboard. Our models outperform prior results and set a new state-of-the-art – ViLBERT with CC + VQA pretraining on MRR, R@k, MR metrics, and further finetuning with a CE loss on dense annotations on NDCG. Finally, adding NSP loss along with CE (as in Section 4.3) offers a balance between optimizing metrics that reward both sparse (original ground-truth answers) and dense annotations.

6 Analysis

As described in Section 5, finetuning on dense annotations leads to a significant increase in NDCG, but hurts the other 5 metrics – MRR, R@1, R@5, R@10 and MR – which depend on the original sparse annotations in VisDial *i.e.* follow-up answers provided in human-human dialog.

We begin by visualizing the distribution of dense relevance scores for these sparse ground-truth (GT) answers in Fig. 2b and observe that $\sim 50\%$ GT answers have relevance ≤ 0.8 , and $\sim 30\%$ have relevance ≤ 0.6 . Thus, there is some degree of

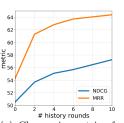
⁵ evalai.cloudcv.org/web/challenges/challenge-page/161/leaderboard/483

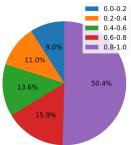
Table 4: Results on VisDial v1.0 test-std. \uparrow indicates higher is better. \downarrow indicates lower
is better. \dagger denotes ensembles. Best single-model results are ${\bf bolded}$ and best ensemble
results are <u>underlined</u> . \star denotes the winning team of the 2019 Visual Dialog Challenge.

$ \begin{array}{c} {} {} {} {} {} {} {} {} {} {} {} {} {}$
$ \begin{array}{c c c c c c c c c c c c c c c c c c c $
$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$
$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$
$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$
$\begin{tabular}{ c c c c c c c c c c c c c c c c c c c$
LF 45.31 55.42 40.95 72.45 82.83 5.9
$ \begin{array}{c} {}_{3}{}_{1}\\ {}_{4}\\ {}_{4}\\ {}_{5}\\ {}_{4}\\ {}_{5}\\ {}_{4}\\ {}_{5}\\ {}_{4}\\ {}_{5}\\ {}_{4}\\ {}_{5}\\ $
$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$
$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$
$ \begin{array}{c} \overbrace{5}\\ \overbrace{5}\\ \overbrace{5}\\ \overbrace{5}\\ \hline{5}\\ \hline\hline \hline{5}\\ \hline\hline \hline{5}\\ \hline\hline \hline $
$\begin{array}{c ccccccccccccccccccccccccccccccccccc$
ि UET-VNU 57.40 59.50 45.50 76.33 85.82 5.3
$\ddot{\dashv}$ square 60.16 61.26 47.15 78.73 88.48 4.4
MS D365 AI 64.47 53.73 42.45 64.68 75.68 6.6
(Random init 60.40 65.53 51.03 83.45 91.83 3.6
$ \begin{array}{c} \begin{array}{c} & \\ & \\ & \\ & \\ \end{array} \end{array} \begin{array}{c} & \\ & \\ & \\ & \\ \end{array} \begin{array}{c} \\ & \\ \\ & \\ \end{array} \begin{array}{c} \\ \\ & \\ \end{array} \begin{array}{c} \\ & \\ \\ & \\ \end{array} \begin{array}{c} \\ \\ & \\ \end{array} \begin{array}{c} \\ & \\ \end{array} \begin{array}{c} \\ \\ \end{array} \begin{array}{c} \\ \\ \\ \end{array} \begin{array}{c} \\ \end{array} \end{array}{c} \end{array} \begin{array}{c} \\ \end{array} \begin{array}{c} \\ \end{array} \begin{array}{c} \\ \end{array} \end{array}{c} \end{array} \begin{array}{c} \\ \end{array} \begin{array}{c} \\ \end{array} \end{array}{c} \end{array} \begin{array}{c} \\ \end{array} \begin{array}{c} \\ \end{array} \end{array}{c} \end{array} \end{array} \begin{array}{c} \\ \end{array} \end{array}{c} \end{array} \end{array} \begin{array}{c} \\ \end{array} \end{array}{c} \end{array} \end{array} \begin{array}{c} \\ \end{array} \end{array} \end{array} \begin{array}{c} \\ \end{array} \end{array}{c} \end{array} \end{array} \end{array} \begin{array}{c} \\ \end{array} \end{array}{c} \end{array} \end{array} \begin{array}{c} \\ \end{array} \end{array} \end{array} \end{array} \end{array} \begin{array}{c} \\ \end{array} \end{array} \end{array} \end{array} \end{array} \end{array} \begin{array}{c} \\ \end{array} $
• CE 74.47 50.74 37.95 64.13 80.00 6.2
$CE + NSP \qquad 68.08 63.92 50.78 79.53 89.60 4.2$

misalignment between dense and sparse annotations – answers originally provided during human-human dialog in VisDial were not always judged to be relevant by all humans during the post-hoc dense annotation phase.

Why are GT and dense annotations misaligned? We notice that many questions with discrepancy between GT and dense annotations are somewhat subjective. For e.g., in row 1, round 7 (Fig. 5), Q: 'what color is the chair?', the GT answer is 'black' but the chair is in shadow and it is difficult to accurately identify its color. And thus, we expect to see variance when multiple humans are polled for the answer. Instead, the GT answer is just one sample from the human answer distribution, not necessarily from its peak. In general, the dense





(a) Change in metrics for varying number of rounds in dialog history for the language-only model trained with NSP + LM

(b) Distribution of dense annotation relevance scores for ground-truth answers in VisDial v1.0 val. $\sim 50\%$ ground-truth answers have relevance scores < 0.8, and $\sim 10\%$ have scores < 0.2 Fig. 2

Relevance Score w/ CC $+$	VQA	CE C	E + NSP
0.0 - 0.2	6.47	14.88	10.79
0.2 - 0.4	4.77	11.11	6.62
0.4 - 0.6	4.02	8.49	4.86
0.6 - 0.8	3.12	6.63	3.77
0.8 - 1.0	1.95	3.26	2.21

11

(c) Mean rank (lower is better) of the GT answers on VisDial v1.0 val split across model variants and ranges of relevance scores

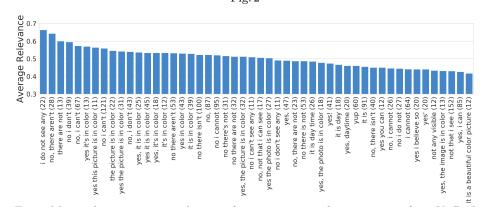


Fig. 3: Mean relevance scores and counts for top-50 most-relevant answers from VisDial v1.0 val dense annotations. These contain several sets of paraphrases – {"yes it's in color", "yes this picture is in color", "the picture is in color", "yes the picture is in color", "yes, it is in color", "yes, it is in color", "yes, it is in color", "yes, it's in color", "yes in color", *etc.* and have a bias towards binary answers

annotations seem less wrong than GT (as they are sourced by consensus) since they are safer – often resolving to answers like 'I cannot tell' when there is uncertainty / subjectivity – but also uninformative – not conveying additional information e.g. 'I think 3 but they are occluded so it is hard to tell' – since such nuanced answers are not part of the list of answer options in VisDial [1].

Model performance on GT vs. dense annotations. Table 2c shows mean ranks of these GT answers as predicted by three model variants – ViLBERT w/ CC + VQA, CE, and CE + NSP – grouped by dense relevance scores. The 'CE' model gets worse mean ranks than 'w/ CC + VQA' for all GT answers, since it is no longer trained with these GT answers during dense annotation finetuning. The CE model assigns low mean ranks to GT answers with higher relevance scores (≥ 0.8), which translates to a high NDCG score (Table 3). But it assigns

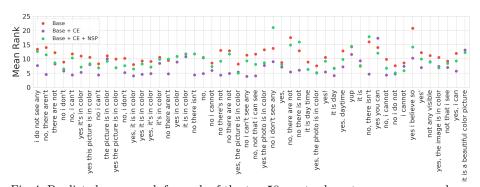


Fig. 4: Predicted mean rank for each of the top-50 most relevant answers as per dense annotations (from Fig. 3) by three model variants – ViLBERT w/ CC + VQA (called 'Base'), CE, and CE + NSP. The CE model gets lower mean ranks for most answers in this set compared to Base. This leads to significantly higher NDCG, as reported in Table 3 and Table 4, but low MRR, since these relevant answers as per dense annotations do not correlate well with the set of original ground-truth answers, as shown in Fig. 2b

poor mean ranks to GT answers with relatively lower relevance scores (≤ 0.8), and since $\sim 50\%$ GT answers have relevance scores ≤ 0.8 , this hurts MRR, R@k, MR for the CE model (Table 3).

Next, we consider the top-50 most-relevant answer options (occurring ≥ 10 times) as per dense annotations in VisDial v1.0 val (not restricting ourselves to only GT answers). Fig. 3 shows the mean relevance scores for this set, and Fig. 4 shows the mean ranks assigned to these answers by our models. The CE model gets better mean ranks in this set compared to Base, leading to high NDCG.

Qualitative examples. Finally, we present uniformly sampled example answer predictions on VisDial v1.0 val from our models along with the ground-truth dialog sequences in Fig. 5 and present additional samples in the appendix. In these examples, consistent with the Visual Dialog task definition [1], at every round of dialog, the model gets the image, ground-truth human dialog history (including caption), and follow-up question as input, and predicts the answer. Specifically, the model ranks 100 answer options. Here we show the top-1 prediction.

We make a few observations. 1) The Base model is surprisingly accurate, e.g. in row 2, round 1 (Fig. 5), Q: 'can you see any people?', predicted answer: 'part of a person', in row 2, round 10, Q: 'anything else interesting about the photo?', predicted answer: 'the dog is looking up at the person with his tongue out'. 2) The CE model often answers with generic responses (such as 'I cannot tell'), especially for questions involving some amount of subjectivity / uncertainty, e.g.in row 1, round 7, Q: 'what color is the chair?', predicted answer: 'I cannot tell' (the chair seems to be in shadow in the image), in row 2, round 7, Q: 'does the dog look happy?', predicted answer: 'I can't tell' (subjective question). 3) This also highlights a consequence of misalignment between ground-truth and dense annotations. While the ground-truth answer provides *one* reasonable response for the question asked, it is answerer-specific to quite an extent and there may

Image + Caption	GT Dialog	Base Model	Base Model + CE	Base Model + CE + NSP
- •	Q: are there any people?	Q: are there any people?	Q: are there any people?	Q: are there any people?
	A: yes	A: no	A: no	A: no
	Q: is the picture in color? A: ves	Q: is the picture in color? A: ves	Q: is the picture in color? A: yes, it is	Q: is the picture in color? A: yes
	Q: how many people are there?	Q: how many people are there?	Q: how many people are there?	Q: how many people are there?
11	A: i can't tell Q: what are they doing?	A: 1 Q: what are they doing?	A: 0 Q: what are they doing?	A: 1 Q: what are they doing?
	A: sitting on the boats	A: sitting on the boats	A: i cannot tell	A: sitting on the boats
and the second s	Q: is this near a beach?	Q: is this near a beach?	Q: is this near a beach?	Q: is this near a beach?
11/2	A: yes Q: are they big boats?	A: yes Q: are they big boats?	A: yes Q: are they big boats?	A: yes Q: are they big boats?
The second se	A: 1 is	A: no	A: nope	A: no
a chair sitting on the beach	Q: what color is the chair? A: white	Q: what color is the chair? A: black	Q: what color is the chair? A: i cannot tell	Q: what color is the chair? A: black
with boats in view offshore	Q: what kind of chair is it?	Q: what kind of chair is it?	Q: what kind of chair is it?	Q: what kind of chair is it?
	A: beach	A: beach	A: i can't tell	A: beach
	Q: what is it made out of? A: wood	Q: what is it made out of? A: metal	Q: what is it made out of? A: i can't tell	Q: what is it made out of? A: i can't tell
	Q: is it a nice day?	Q: is it a nice day?	Q: is it a nice day?	Q: is it a nice day?
	A: overcast	A: yes	A: yes	A: yes
	Q: can you see any people?	Q: can you see any people?	Q: can you see any people?	Q: can you see any people?
	A: part of a person Q: is the photo in color?	A: part of a person	A: yes	A: yes Q: is the photo in color?
	A: yes	Q: is the photo in color? A: yes	Q: is the photo in color? A: ves	A: yes
	Q: man or woman? A: i can't tell	Q: man or woman?	Q: man or woman?	Q: man or woman? A: can't tell
	Q: how many dogs?	A: can't see his face	A: can't tell	Q: how many dogs?
	A: just 1	Q: how many dogs? A: 1	Q: how many dogs? A: 1	A: 1
	Q: is the dog big or small? A: medium	Q: is the dog big or small?	Q: is the dog big or small?	Q: is the dog big or small? A: it looks like a large dog
	Q: inside or outside?	A: big dog O: inside or outside?	A: big dog O: inside or outside?	Q: inside or outside?
	A: hard to tell but i think outside Q: does the dog look happy?	A: outside	A: outside	A: outside O: does the dog look happy?
AT IN THE	A: kind of	Q: does the dog look happy? A: ves	Q: does the dog look happy? A: i can't tell	A: yes
a brown and white dog riding a skateboard	Q: can you see the ground?	Q: can you see the ground?	Q: can you see the ground?	Q: can you see the ground?
skateboard	A: yes Q: what color is the skateboard?	A: yes	A: yes	A: yes Q: what color is the skateboard?
	A: black	Q: what color is the skateboard? A: black	Q: what color is the skateboard? A: black	A: black
	Q: anything else interesting about the photo?	Q: anything else interesting about the photo?	Q: anything else interesting about	Q: anything else interesting about the photo?
	A: the dog is looking up at the person	A: the dog is looking up at the person with his tongue out	the photo? A: no	A: the dog is looking up at the person
	with his tongue out	tongue out	A. 110	with his tongue out
	Q: how many giraffes are there?		Q: how many giraffes are there?	Q: how many giraffes are there?
	A: 3 Q: are they all adults?	Q: how many giraffes are there? A: there are 4 giraffes	A: 3 O: are they all adults?	A: 3 O: are they all adults?
	A: yes	Q: are they all adults?	A: yes	A: yes
	Q: do you see grass?	A: yes Q: do you see grass?	Q: do you see grass?	Q: do you see grass?
	A: yes Q: are there any trees?	A: yes	A: yes Q: are there any trees?	A: yes Q: are there any trees?
A.	A: yes	Q: are there any trees?	A: yes	A: yes
	Q: is it sunny? A: ves	A: yes O: is it sunny?	Q: is it sunny? A: ves	Q: is it sunny? A: ves
and the season	Q: do the trees have any leaves on	A: yes	Q: do the trees have any leaves on	Q: do the trees have any leaves on
M ATR.	them? A: yes	Q: do the trees have any leaves on them? A: yes	them? A: yes	them? A: yes
	Q: are there any other animals	O: are there any other animals visible?	Q: are there any other animals	Q: are there any other animals
a group of giraffes is standing in a savannah	visible? A: no	A: no Q: are the giraffes heads up?	visible? A: no	visible? A: no
	Q: are the giraffes heads up?	A: yes	Q: are the giraffes heads up?	Q: are the giraffes heads up?
	A: yes	Q: is the sky clear?	A: yes they are	A: yes
	Q: is the sky clear? A: yes	A: yes Q: does it look dry?	Q: is the sky clear? A: yes	Q: is the sky clear? A: yes
	Q: does it look dry?	A: yes	Q: does it look dry?	Q: does it look dry?
	A: yes		A: yes	A: yes
	Q: is the cow large or small?		Q: is the cow large or small?	Ou is the same large as smaller
	A: large		A: large Q: does the cow have an ear tag?	Q: is the cow large or small? A: large
		Q: is the cow large or small?		
	Q: does the cow have an ear tag? A: no	Q: is the cow large or small? A: large	A: yes	Q: does the cow have an ear tag?
	A: no Q: what color is the building?	A: large Q: does the cow have an ear tag?		A: yes
	A: no Q: what color is the building? A: white	A: large Q: does the cow have an ear tag? A: yes Q: what color is the building?	A: yes Q: what color is the building? A: white Q: is the photo professional?	A: yes Q: what color is the building? A: white
	A: no Q: what color is the building? A: white Q: is the photo professional? A: no	A: large Q: does the cow have an ear tag? A: yes Q: what color is the building? A: white	A: yes Q: what color is the building? A: white Q: is the photo professional? A: nope	A: yes Q: what color is the building? A: white Q: is the photo professional?
	A: no Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well	A: large Q: does the cow have an ear tag? A: yes Q: what color is the building? A: white Q: is the photo professional? A: no	A: yes Q: what color is the building? A: white Q: is the photo professional?	A: yes Q: what color is the building? A: white Q: is the photo professional? A: no
	A: no Q: what color is the building? A: white Q: is the photo professional? A: no	A: large Q: does the cow have an ear tag? A: yes Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed?	A: yes Q: what color is the building? A: white Q: is the photo professional? A: nope Q: does the grass look well groomed? A: yes	A: yes Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed' A: yes
	A: no Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed? A: yes Q: are there any trees in the photo?	A: large Q: does the cow have an ear tag? A: yes Q: what color is the building? A: white Q: is the photo professional? A: no	A: yes Q: what color is the building? A: white Q: is the photo professional? A: nope Q: does the grass look well groomed?	A: yes Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed'
	A: no Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed? A: yes Q: are there any trees in the photo? A: no	A: large Q: does the cow have an ear tag? A: yes Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed? A: yes Q: are there any trees in the photo? A: no	A: yes Q: what color is the building? A: white Q: is the photo professional? A: nope Q: does the grass look well groomed? A: yes Q: are there any trees in the photo? A: yes Q: are there any fences surrounding Q: are there any fences surrounding	A: yes Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed A: yes Q: are there any trees in the photo? A: yes Q: are there any fences surrounding Q: are there any fences surrounding
the large brown cow is	A: no Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed? A: yes Q: are there any trees in the photo? A: no Q: are there any fences surrounding the cow?	A: large Q: does the cow have an ear tag? A: yes Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed? A: yes Q: are there any trees in the photo? A: are not ear only chees surrounding the cow? P: are there any fences surrounding the cow?	A: yes Q: what color is the building? A: white Q: is the photo professional? A: nope Q: does the grass look well groomed? A: yes Q: are there any trees in the photo? A: yes Q: are there any fences surrounding the cow?	A: yes Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed A: yes Q: are there any trees in the photo? A: yes Q: are there any fences surrounding the cow?
the large brown cow is standing by the building eating arras	A: no Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed? A: yes Q: are there any trees in the photo? A: no Q: are there any fences surrounding the cow? A: no	A: large Q; Q: does the cow have an ear tag? A; A: yes Q; what color is the building? A: white Q; is the photo professional? A: no Q; does the grass look well groomed? Q; does the grass look well groomed? Q; Q; arce there any trees in the photo? A; no Q; arce there any fences surrounding the cow? Q; Q; alocs is look sunny? Q;	A: yes Q: what color is the building? A: white Q: is the photo professional? A: nope Q: does the grass look well groomed? A: yes Q: are there any trees in the photo? A: yes Q: are there any fences surrounding the cow? Q: does is look sunny?	A: yes Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed A: yes Q: are there any trees in the photo? A: yes Q: are there any fences surrounding Q: are there any fences surrounding
standing by the building eating	A: no Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed? A: yes Q: are there any trees in the photo? A: no Q: are there any fences surrounding the cow? A: no Q: does is look sunny? A: yes	A: large Q: does the cow have an ear tag? A: yes Q: shat color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed? A: yes Q: are there any trees in the photo? A: no Q: are there any fences surrounding the cow? A: no Q: ace is look sunny? A: yes	A: yes Q: what color is the building? A: white Q: is the photo professional? A: nope Q: does the grass look well groomed? A: yes Q: arc there any trees in the photo? A: yes Q: arc there any fences surrounding the cow? A: no it does not	A: yes Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed? A: yes Q: are there any trees in the photo? A: yes Q: are there any fences surrounding the cow? A: no Q: does is look sunny? A: no
standing by the building eating	A: no Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed? A: yes Q: are there any teres in the photo? A: no Q: are there any fences surrounding the cow? A: no Q: are there any fences surrounding the cow? A: no Q: does is look sunny? A: yes Q: does us look sunny? A: yes Q: can you see the sky in the picture?	A: large Q; Q: does the cow have an ear tag? A; A: yes Q; what color is the building? A: white Q; is the photo professional? A: no Q; does the grass look well groomed? Q: does the grass look well groomed? Q; are there any trees in the photo? A: no Q; are there any frees surrounding the cow? Q: are there any focus surrounding the cow? Q; ale so look sunny?	A: yes Q: what color is the building? A: white Q: is the photo professional? A: nope Q: does the grass look well groomed? A: yes Q: are there any trees in the photo? A: yes Q: are there any fences surrounding the cow? Q: does is look sunny? Q: does is look sunny? A: no i Q: does not does not O: can you see the sky in the	A: yes Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed? A: yes Q: are there any trees in the photo? A: yes Q: are there any fences surrounding the cow? A: no Q: does is look sunny? A: no
standing by the building eating	A: no Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed? A: yes Q: are there any trees in the photo? Q: are there any fences surrounding the cow? A: no Q: does is look sunny? A: yes Q: can you see the sky in the picture? A: no	A: large Q: does the cow have an ear tag? A: yes Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed? A: yes Q: are there any trees in the photo? A: no Q: are there any trees in the photo? A: no Q: are there any tenes surrounding the cow? Q: does its look sunny? A: yes Q: can you see the sky in the picture? A: no Q: can you see windows in the building?	A: yes Q: what color is the building? A: white Q: is the photo professional? A: nope Q: does the grass look well groomed? A: yes Q: are there any trees in the photo? A: yes Q: are there any fences surrounding the cow? Q: does is look sunny? A: no i Q: does is look sunny? A: no it does not Q: can you see the sky in the picture?	A: yes Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed: A: yes Q: are there any trees in the photo? A: yes Q: are there any fences surrounding the cow? A: no Q: does is look sunny? A: no Q: can you see the sky in the picture? A: no Q: can you see windows in the
standing by the building eating	A: no Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed? A: yes Q: are there any teres in the photo? A: no Q: are there any fences surrounding the cow? A: no Q: are there any fences surrounding the cow? A: no Q: does is look sunny? A: yes Q: does us look sunny? A: yes Q: can you see the sky in the picture?	A: large Q: does the cow have an ear tag? A: yes Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed? A: yes Q: ano there any trees in the photo? Q: are there any fences surrounding the cow? A: no Q: are there any fences surrounding the cow? A: no Q: does is look sunny? A: yes Q: can you see the sky in the picture? A: no	A: yes Q: what color is the building? A: white Q: is the photo professional? A: nope Q: does the grass look well growned? A: yes Q: are there any trees in the photo? A: yes Q: are there any fences surrounding the cow? A: no it does not Q: does is look sunny? A: no it does not Q: any you see the sky in the picture?	A: yes Q: what color is the building? A: white Q: is the photo professional? A: no Q: does the grass look well groomed A: yes Q: are there any trees in the photo? A: yes Q: are there any fences surrounding the cow? A: no Q: does is look sunny? A: no Q: any you see the sky in the picture? A: no

Fig. 5: Qualitative samples for three model variants – ViLBERT w/ CC + VQA (called 'Base'), Base + CE, and Base + CE + NSP

be other correct answers (annotated in the dense annotations). A negative effect of this misalignment is that when finetuned on dense annotations (CE), the model gets rewarded for generic answers (e.g. 'cannot tell'). While being able

to capture and reason about uncertainty is a desirable property models should have, it would be more helpful if these agents can convey more information with appropriate qualifiers (*e.g.* 'I think 3 but they are occluded so it is hard to tell') than a blanket 'I cannot tell'. We aim to study this in future work.

7 Implementation

We use the BERT_{BASE} model [36] for the linguistic stream. We use 6 layers of Transformer blocks (with 8 attention heads and a hidden state size of 1024) for the visual stream. The co-attention layers connect the 6 Transformer layers in the visual stream to the last 6 Transformer layers in the linguistic stream. We train on dialog sequences with atmost 256 tokens as most sequences had atmost 256 tokens. During inference, we truncate longer sequences by removing rounds starting from round 1 (we keep the caption). We set all loss coefficients to 1. We use a batch size of 128 for language-only experiments and 80 for other experiments. We use Adam [63] and linearly increase learning rate from 0 to $2e^{-5}$ over 10k iterations and decay to $1e^{-5}$ over 200k iterations. Our code is available at github.com/vmurahari3/visdial-bert/.

8 Conclusion

We introduce a model for Visual Dialog that enables pretraining on large-scale image-text datasets before transferring and finetuning on VisDial. Our model is an adaptation of ViLBERT [2], and our best single model is pretrained on BooksCorpus [30], English Wikipedia (at the BERT stage), and on Conceptual Captions [3], VQA [4] (at the ViLBERT stage), before finetuning on VisDial, optionally with dense annotations. Our model outperforms prior published results by > 1% absolute on NDCG and MRR, achieving state-of-the-art results, and providing a simple baseline for future 'pretrain-then-transfer' approaches.

Through careful analysis of our results, we find that the recently released dense annotations for the task do not correlate well with the original ground-truth dialog answers, leading to a trade-off when models optimize for metrics that take into account these dense annotations (NDCG) vs. the original sparse annotations (MRR). This opens up avenues for future research into better evaluation metrics.

Finally, note that our model is discriminative – it can pick a good answer from a list of answer options – but cannot generate an answer. In the future, we aim to develop robust decoding techniques, based on decoding strategies for transformer-based models introduced in [34, 64], for a strong generative model.

Acknowledgments. The Georgia Tech effort was supported in part by NSF, AFRL, DARPA, ONR YIPs, ARO PECASE, Amazon. AD was supported by fellowships from Facebook, Adobe, Snap Inc. Views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government, or any sponsor.

References

- A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra, "Visual Dialog," in *CVPR*, 2017. 1, 2, 3, 4, 5, 6, 8, 11, 12
- J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeurIPS*, 2019. 1, 2, 3, 4, 5, 14
- P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in ACL, 2018.
 1, 2, 3, 4, 5, 7, 9, 10, 14
- 4. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual Question Answering," in *ICCV*, 2015. 1, 2, 3, 4, 5, 7, 9, 10, 14
- H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville, "GuessWhat?! visual object discovery through multi-modal dialogue," in *CVPR*, 2017. 1, 3, 4
- F. Strub, H. De Vries, J. Mary, B. Piot, A. Courville, and O. Pietquin, "End-toend optimization of goal-driven and visually grounded dialogue systems," arXiv preprint arXiv:1703.05423, 2017. 1, 3, 4
- A. Das, S. Kottur, J. M. Moura, S. Lee, and D. Batra, "Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning," in *ICCV*, 2017. 1, 3, 4
- J. Lu, A. Kannan, J. Yang, D. Parikh, and D. Batra, "Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model," in NIPS, 2017. 1, 3, 4
- D. Massiceti, N. Siddharth, P. K. Dokania, and P. H. Torr, "Flipdial: A generative model for two-way visual dialogue," in CVPR, 2018. 1, 3, 4
- Q. Wu, P. Wang, C. Shen, I. Reid, and A. van den Hengel, "Are you talking to me? reasoned visual dialog generation through adversarial learning," in *CVPR*, 2018. 1, 3, 4
- 11. U. Jain, S. Lazebnik, and A. G. Schwing, "Two can play this game: visual dialog with discriminative question generation and answering," in *CVPR*, 2018. 1, 3
- S. Kottur, J. M. Moura, D. Parikh, D. Batra, and M. Rohrbach, "Visual coreference resolution in visual dialog using neural module networks," in *ECCV*, 2018. 1, 3, 4, 10
- 13. S.-W. Lee, T. Gao, S. Yang, J. Yoo, and J.-W. Ha, "Large-scale answerer in questioner's mind for visual dialog question generation," in *ICLR*, 2019. 1, 3, 4
- 14. Y. Niu, H. Zhang, M. Zhang, J. Zhang, Z. Lu, and J.-R. Wen, "Recursive visual attention in visual dialog," in *CVPR*, 2019. 1, 3, 4, 10
- 15. Z. Zheng, W. Wang, S. Qi, and S.-C. Zhu, "Reasoning visual dialogs with structural and partial observations," in *CVPR*, 2019. 1, 3, 10
- 16. I. Schwartz, S. Yu, T. Hazan, and A. G. Schwing, "Factor graph attention," in *CVPR*, 2019. 1, 3, 10
- 17. G.-C. Kang, J. Lim, and B.-T. Zhang, "Dual attention networks for visual reference resolution in visual dialog," in *EMNLP*, 2019. 1, 3, 10
- Z. Gan, Y. Cheng, A. E. Kholy, L. Li, J. Liu, and J. Gao, "Multi-step reasoning via recurrent dual attention for visual dialog," in ACL, 2019. 1, 3, 10
- S. Kottur, J. M. Moura, D. Parikh, D. Batra, and M. Rohrbach, "Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog," in NAACL, 2019. 1, 3
- 20. V. Murahari, P. Chattopadhyay, D. Batra, D. Parikh, and A. Das, "Improving generative visual dialog by answering diverse questions," in *EMNLP*, 2019. 1, 3, 4

Vishvak Murahari et al.

- R. Shekhar, A. Venkatesh, T. Baumgärtner, E. Bruni, B. Plank, R. Bernardi, and R. Fernández, "Beyond task success: A closer look at jointly learning to see, ask, and guesswhat," in NAACL, 2019. 1, 3
- 22. T. Yang, Z.-J. Zha, and H. Zhang, "Making history matter: Gold-critic sequence training for visual dialog," *arXiv preprint arXiv:1902.09326*, 2019. 1, 3, 10
- D. Guo, C. Xu, and D. Tao, "Image-question-answer synergistic network for visual dialog," in CVPR, 2019. 1, 3, 10
- J. Qi, Y. Niu, J. Huang, and H. Zhang, "Two causal principles for improving visual dialog," arXiv preprint arXiv:1911.10496, 2019. 1, 3, 10
- X. Jiang, J. Yu, Z. Qin, Y. Zhuang, X. Zhang, Y. Hu, and Q. Wu, "DualVD: An adaptive dual encoding model for deep visual understanding in visual dialogue," in AAAI, 2020. 1, 3, 10
- H. Alamri, V. Cartillier, A. Das, J. Wang, A. Cherian, I. Essa, D. Batra, T. K. Marks, C. Hori, P. Anderson, S. Lee, and D. Parikh, "Audio visual scene-aware dialog," in *CVPR*, 2019.
- 27. H. de Vries, K. Shuster, D. Batra, D. Parikh, J. Weston, and D. Kiela, "Talk the walk: Navigating new york city through grounded dialogue," arXiv preprint arXiv:1807.03367, 2018. 1
- 28. K. Nguyen and H. Daumé III, "Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning," in EMNLP, 2019. 1
- 29. J. Thomason, M. Murray, M. Cakmak, and L. Zettlemoyer, "Vision-and-dialog navigation," 2019. 1
- 30. Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *ICCV*, 2015. 2, 7, 8, 14
- K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in CVPR, 2016. 2, 3, 5
- 32. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015. 2, 3
- 33. A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *NIPS*, 2012. 2, 3
- A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding with unsupervised learning," 2018. 2, 3, 4, 14
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017. 2, 3, 4, 7
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in NAACL, 2019. 2, 3, 4, 5, 7, 14
- 37. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019. 2, 3, 4
- Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "AL-BERT: A lite BERT for self-supervised learning of language representations," arXiv preprint arXiv:1909.11942, 2019. 2, 3, 4
- Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," arXiv preprint arXiv:1906.08237, 2019. 2, 3, 4
- 40. C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," arXiv preprint arXiv:1910.10683, 2019. 2, 3, 4

- 41. Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "DialoGPT: Large-scale generative pre-training for conversational response generation," arXiv preprint arXiv:1911.00536, 2019. 2, 3, 4
- O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *IJCV*, 2015. 2, 4
- 43. R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, 2017. 2, 5
- 44. A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," arXiv preprint arXiv:1804.07461, 2018. 2, 4
- 45. L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A simple and performant baseline for vision and language," arXiv preprint arXiv:1908.03557, 2019. 2, 4
- 46. H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," arXiv preprint arXiv:1908.07490, 2019. 2, 4
- 47. Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "UNITER: Learning UNiversal Image-TExt Representations," arXiv preprint arXiv:1909.11740, 2019. 2, 4
- 48. G. Li, N. Duan, Y. Fang, D. Jiang, and M. Zhou, "Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training," arXiv preprint arXiv:1908.06066, 2019. 2, 4
- W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "VL-BERT: Pre-training of generic visual-linguistic representations," arXiv preprint arXiv:1908.08530, 2019.
 2, 4
- 50. C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," arXiv preprint arXiv:1904.01766, 2019. 2, 4
- 51. R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *CVPR*, 2019. 2, 4
- 52. A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi, "A corpus for reasoning about natural language grounded in photographs," in *ACL*, 2019. 2, 4
- 53. N. Xie, F. Lai, D. Doran, and A. Kadav, "Visual entailment: A novel task for fine-grained image understanding," arXiv preprint arXiv:1901.06706, 2019. 2, 4
- P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *TACL*, 2014. 2, 4
- 55. K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for imagetext matching," in *ECCV*, 2018. 2, 4
- S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg, "ReferItGame: Referring to Objects in Photographs of Natural Scenes," in *EMNLP*, 2014. 2, 4
- 57. W. Hao, C. Li, X. Li, L. Carin, and J. Gao, "Towards learning a generic agent for vision-and-language navigation via pre-training," in *CVPR*, 2020. 2, 4
- A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), pp. 8024–8035, Curran Associates, Inc., 2019. 3

Vishvak Murahari et al.

- T. Wolf, V. Sanh, J. Chaumond, and C. Delangue, "TransferTransfo: A transfer learning approach for neural network based conversational agents," arXiv preprint arXiv:1901.08149, 2019. 5
- P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," arXiv preprint arXiv:1707.07998, 2017. 5
- S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015. 5
- T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. DollÄ_ir, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *ECCV*, 2014.
- 63. D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *ICLR*, 2015. 14
- 64. M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," arXiv preprint arXiv:1910.13461, 2019. 14

9 Appendix

9.1 Negative Results

To encourage the model to learn sentence level semantics, we tried a pretraining strategy which we refer to as the inconsistency loss. During pretraining, we randomly select answers at 3 different rounds and replace them randomly with one of the 100 answer options at those rounds. Similar to the masked language modeling loss, at each token, we predict the probability of the token being "inconsistent" in the dialog history and we assume that all the tokens in the randomly selected answer option are "inconsistent". We also only consider dialog sequences which have at least 6 rounds for creating samples. This is to make sure that the model has enough context to make accurate predictions. We hoped that this loss would encourage the model to capture sentence level semantics as the model would need to figure out which sentences fit in together in a dialog sequence.

A pitfall in this setting is that some of the 100 answer options at each round are often similar to the GT answer or are generic responses (*e.g.* "yes", "no", "maybe", *etc.*). Thus, there is a chance that swapping the GT answer with a randomly selected answer option might lead to a consistent dialog sequence. We instead try to create a new sample by randomly selecting a round and reordering/jumbling the answers at that round and the answers at the preceding and the following rounds. We hope that jumbling the order of answers would lead to an inconsistent dialog sequence. We call this variant "inconsistency loss (jumbled)".

We cannot use the batch of data used to calculate the NSP and MLM loss to calculate the inconsistency loss. We first try to use half the batch to calculate the inconsistency loss and the other half to calculate the NSP and MLM losses. We then try to calculate the inconsistency loss and the NSP and MLM losses in alternating batches. We present results for models trained with multiple variants of the inconsistency loss in the language-only setting in Table 5.

We do not see any significant improvement by training with the inconsistency loss. We note that creating samples by reordering answers in different rounds does not improve performance. We also note that optimizing for the inconsistency loss and the NSP and MLM losses in alternating batches leads to improvements.

Table 5: Performance of language-only models on VisDial v1.0 val, trained with different sets of pretraining losses including the inconsistency loss.

Variant	NDCG \uparrow	MRR ↑	$R@1\uparrow$	R@5 \uparrow	$R@10 \uparrow$	$\mathrm{MR}\downarrow$
NSP + MLM	57.22	64.10	50.05	81.09	90.00	4.16
NSP + MLM + Inconsistency	56.22	64.13	49.88	81.35	90.11	4.01
NSP + Inconsistency	55.66	63.00	48.63	80.61	89.36	4.20
NSP + Inconsistency (jumbled)	55.83	63.08	48.71	80.47	89.49	4.19
NSP + MLM + Inconsistency (alternating batches)	57.17	64.21	50.04	81.48	90.02	4.03

9.2 Additional results: Dense annotation finetuning

We vary the coefficients for CE and NSP losses during dense annotation finetuning (α for CE loss and $1 - \alpha$ for NSP loss) and report results in Table 6. α acts as a lever balancing metrics based on dense annotations (NDCG) and metrics based on sparse annotations (MRR, R@1, R@5, R@10).

Table 6: Performance on VisDial v1.0 val for variants finetuned on dense annotations with a coefficient of α for the CE loss and $1 - \alpha$ for the NSP loss. α balances NDCG and MRR performance with higher α leading to higher NDCG and lower MRR.

α	NDCG \uparrow	$MRR \uparrow$	R@1 \uparrow	R@5 \uparrow	$R@10 \uparrow$	$\mathrm{MR}\downarrow$
0.25	68.29	65.96	53.34	81.42	90.37	4.12
0.5	69.04	65.58	53.04	80.64	90.01	4.40
0.75	71.17	62.88	49.49	78.91	88.83	4.39
1.0 (CE)	75.10	52.12	39.84	64.93	80.47	6.26

9.3 Additional Qualitative Samples

We present some additional qualitative samples in Fig. 6 and Fig. 7.

Image + Caption	GT Dialog	Base Model	Base Model + CE	Base Model + CE + NSP
	Q: is the light lit up?	Q: is the light lit up?	Q: is the light lit up?	Q: is the light lit up?
	A: yes	A: yes	A: yes	A: yes
	Q: are there cars? A: yes	Q: are there cars? A: yes	Q: are there cars? A: yes	Q: are there cars? A: yes
	Q: is this a city area?	O: is this a city area?	Q: is this a city area?	Q: is this a city area?
A showing a	A: maybe , i'm not sure	A: yes	A: yes	A: yes
	Q: is there people?	Q: is there people?	Q: is there people?	Q: is there people?
	A: no Q: is it daytime?	A: no Q: is it daytime?	A: no Q: is it daytime?	A: yes, at least 1 Q: is it daytime?
	A: no	A: no	A: yes	A: yes
12-11-11-15-15-15-	Q: is there a street sign?	Q: is there a street sign?	Q: is there a street sign?	Q: is there a street sign?
Contraction of the second s	A: no	A: no	A: yes	A: yes
e intersection had a red light	Q: is it dark? A: yes	Q: is it dark? A: yes	Q: is it dark? A: yes	Q: is it dark?
d wet streets	Q: is the light metal?	O: is the light metal?	Q: is the light metal?	A: yes Q: is the light metal?
	A: yes	A: yes	A: no	A: yes
	Q: can you see buildings?	Q: can you see buildings?	Q: can you see buildings?	Q: can you see buildings?
	A: yes Q: can you see any trees?	A: yes Q: can you see any trees?	A: yes Q: can you see any trees?	A: yes Q: can you see any trees?
	A: yes	A: yes	A: yes	A: yes
	-			
	Q: are there any people? A: no	Q: are there any people? A: no	Q: are there any people? A: no	Q: are there any people? A: no
	Q: what color is the cup?	Q: what color is the cup?	Q: what color is the cup?	Q: what color is the cup?
	A: blue	A: black	A: black	A: black
	Q: what color are the toothbrushes?	Q: what color are the toothbrushes?	Q: what color are the toothbrushes?	Q: what color are the toothbrush
	A: red and blue Q: is the sink visible?	A: orange O: is the sink visible?	A: white O: is the sink visible?	A: orange O: is the sink visible?
ALL	A: no	A: no	A: no	A: no
	Q: is a cabinet visible?	Q: is a cabinet visible?	Q: is a cabinet visible?	Q: is a cabinet visible?
	A: no	A: no	A: no	A: no
	Q: are they adult toothbrushes?	Q: are they adult toothbrushes?	Q: are they adult toothbrushes?	Q: are they adult toothbrushes?
	A: they look like it Q: is the ground visible?	A: yes Q: is the ground visible?	A: yes Q: is the ground visible?	A: yes Q: is the ground visible?
toothbrushes sit in a blue cup	A: no	A: no	A: no	A: no
a bathroom	Q: is there a mirror there?	Q: is there a mirror there?	Q: is there a mirror there?	Q: is there a mirror there?
	A: yes, i can see the corner of it	A: no	A: no	A: no
	Q: it a close up?	Q: it a close up?	Q: it a close up? A: yes	Q: it a close up?
	A: yes Q: is the counter visible?	A: yes Q: is the counter visible?	Q: is the counter visible?	A: yes Q: is the counter visible?
	A: no	A: no	A: no, it is not	A: yes
	Q: is the elephant a baby?	Q: is the elephant a baby?	O: is the elephant a baby?	O: is the elephant a baby?
	A: no	A: i think it's probably a young elephant it	A: i think so	A: ves
	Q: is he eating from a tree?	doesn't appear to be very big	Q: is he eating from a tree?	Q: is he eating from a tree?
	A: no the ground	Q: is he eating from a tree?	A: yes	A: yes
	Q: are they outside?	A: yes	Q: are they outside?	Q: are they outside?
ALL AND A	A: yes Q: is there grass?	Q: are they outside? A: yes	A: yes Q: is there grass?	A: yes Q: is there grass?
Start Start	A: yes	Q: is there grass?	A: yes	A: yes
	Q: are there trees around?	A: yes	Q: are there trees around?	Q: are there trees around?
	A: yes	Q: are there trees around?	A: yes	A: yes
Frind William he	Q: is it sunny? A: it is hard to tell	A: yes Q: is it sunny?	Q: is it sunny?	Q: is it sunny?
A SUBJER OF	A: it is hard to tell Q: is the food in his mouth?	Q: is it sunny? A: ves	A: yes Q: is the food in his mouth?	A: yes Q: is the food in his mouth?
elephant eats large amounts	A: yes	Q: is the food in his mouth?	A: yes	A: yes
foliage as another elephant	Q: do the leaves look fresh?	A: yes	Q: do the leaves look fresh?	Q: do the leaves look fresh?
inds nearby	A: yes	Q: do the leaves look fresh?	A: yes	A: yes
	Q: is the other elephant watching him?	A: yes Q: is the other elephant watching him?	Q: is the other elephant watching him?	Q: is the other elephant watching him?
	A: no	A: no	A: i can't tell	A: no
	Q: what color is the grass?	Q: what color is the grass?	Q: what color is the grass?	Q: what color is the grass?
	A: green	A: brown	A: brownish	A: brown and green
	Q: any people?	Q: any people?	Q: any people?	Q: any people?
	A: no	A: no	A: no	A: no
	Q: how many lemons? A: 1	Q: how many lemons? A: 1	Q: how many lemons? A: 2	Q: how many lemons? A: 2
	Q: is it large?	Q: is it large?	Q: is it large?	Q: is it large?
	A: no	A: yes	A: yes	A: no
	Q: is the fruit ripe?	Q: is the fruit ripe?	Q: is the fruit ripe?	Q: is the fruit ripe?
Contraction of the second	A: yes Q: has it been sliced?	A: yes Q: has it been sliced?	A: yes Q: has it been sliced?	A: yes Q: has it been sliced?
A KANA AND	A: no	A: no	A: no	A: no
	Q: is there a plate?	Q: is there a plate?	Q: is there a plate?	Q: is there a plate?
	A: can't tell	A: no	A: no	A: no
emon, peach, and another	Q: what is the other fruit?	Q: what is the other fruit?	Q: what is the other fruit?	Q: what is the other fruit?
ece of fruit all sitting next to	A: apple peach O: is this on a table?	A: apple peach O: is this on a table?	A: apple peach O: is this on a table?	A: apple peach O: is this on a table?
ch other	A: can't tell it is a very close up shot	A: yes	A: can't tell	A: can't tell
	Q: is this in a kitchen?	Q: is this in a kitchen?	Q: is this in a kitchen?	Q: is this in a kitchen?
	A: can't tell	A: no	A: can't tell	A: can't tell
			Q: indoors?	Q: indoors?
	Q: indoors? A: no	Q: indoors? A: yes	A: can't tell	A: yes

Fig. 6: Qualitative samples for three model variants – ViLBERT w/ CC + VQA (called 'Base'), Base + CE, and Base + CE + NSP.

PerspectiveOperation of the second secon					
A: 7: number of the second of the s		Q: how many steps are there?		0.1	0.1
a. b. b. a. b. a. b. b. a. b. a. b. a. b. a. b. a. b. a			Q: how many steps are there?		
a high b high		O: is it a passenger train?			
a. b.			O: is it a passenger train?	Q: is it a passenger train?	
a biask takker 0 - source are insegr? Note insegree in the insegr? Note insegree insegree insegree Note insegree insegree a biask takker 0 - source insegree Note insegree insegree Note insegree insegree Note insegree insegree Note insegree insegree a biask takker Note insegree insegree a biase of a more insegree Note insegree insegree Note insegree insegree Note insegree insegree Note insegree insegree a biase of a more insegree Note insegree insegree Note insegree insegree Note insegree insegree Note insegree insegree a biase of a more insegree Note insegree insegree Note insegree Note insegree Note insegree Note insegree a biase of a more insegree Note insegree Note insegree Note insegree Note insegree Note insegree Note insegree a biase of a more insegree Note insegree Note insegree Note insegree Note insegree Note insegree a biase of a more insegree Note insegree <		O: what color are the steps?	A: yes	A: yes, it is	
a. b Buck Desire there poole on the supp? A Buck Desire there poole on the supp? Desire poole on the supp?			O: what color are the steps?	Q: what color are the steps?	
Image: state of the s		O: are there people on the steps?			
Image: set index is an and back of an investigation of the standard of the	and the second s	A: no	O: are there people on the steps?		
bit		O: can you see windows?			
Bit optimized in controls on the low of the low of the section o					
Bit only set 1, too clear up were the construction of the set of	and it is a loss				
Image: Section of a manage with the section of th	A DECEMBER OF THE OWNER OWNER OF THE OWNER OWNE OWNER OWNE	At only see 1, too close up to see			Q: how many cars are there?
ha das dar angenge, sub sub dasket min car angenge sub dasket min c		A: only see 1, too close up to see		A: just 1	A: 1
his door of manages, building on boots or for manages landing or for landing or for manages landing or for landing or for manages or for manages landing or for landing or for manages or for manages landing or for landing or for manages or for manages landing or for landing or for manages or for manages landing or for manages landing or for manages or manages landing or for manages landing or for manages or manages landing or for manages landing or for manages or manages landing or for manages landing or for manages or manages landing or for manages landing or for manages or manages landing or for manages landing or for manages or manages landing or for manages or manages landing or f				O: are there windows?	O: are there windows?
and black mayer is always to be a set of the second rescalator lock like? A second second rescalator lock					
with days lading up to 11 with days lading up to 11 wi			A: yes	O: what does the weather look like?	O: what does the weather look like?
A brown of the standard of th	with steps leading up to it				
a. Sind of the second of th					
Name Name Name Name Name Name Name a Wardy A whole A whole A whole A whole A whole A whole a Wardy A whole B wh		Q: what can you see through the			window?
e. was color in the sky? Q: what color is the sky? Q: what color is the sky? Q: what color is the sky? a. was in the standard or indoor					
Image: constraint of a line start, and a line of the start, and a line line start, and a line of the start, and a line line					O: what color is the slav?
Image: Constraint of the second of		Q: what color is the sky?	A: white	Q. what color is the sky?	
Image: Second		A: whitish pale blue		A: white	A: blue
Image: Second				1	
a bard of sector balls on any bayes of balls on any bayes? a bard of sector balls on any bayes? a bard of sector balls of any bard any				O: is this an outdoor or indoor	O: is this an outdoor or indoor
A 'you can only see that tables on ian locations in the section of the section of the section of the section of the section is the section i	1				
a hunder Unserte Octos the shelf have many layers? A construction of the sheet have many layers? A construction of the sheet have many layers? A construction of the sheet have many layers? a work of vegetables on a sheet have mark h	1	A: you can only see the table so i am	A: outdoors		
P: does the shelf how rampy logyer? A card use the top P: does the informal multiply by P:	1		Q: does the shelf have many layers?		
A ign 2 that ican see () (P) is the picture in color? (A) is the set in color? (A) is the picture in color? A ign 2 that ican see () (P) is the picture in color? (A) is the picture in color? (A) is the picture in color? A ign 2 that ican see () (P) is the picture in color? (A) is the picture in color? (A) is the picture in color? A ign 2 that ican see () (P) is the picture in color? (A) is the picture in color? (A) is the picture in color? A ign 2 that ican see () (P) is the picture in color? (A) is the picture in color? (A) is the picture in color? A ign 2 that ican see () (P) is the picture in color? (A) is the picture in color? (A) is the picture in color? A ign 2 that ican see () (P) is the picture in color? (A) is the picture in color? (A) is the picture in color? A is the picture in color? (A) is the picture in color? (A) is the picture in color? (A) is the picture in color? A is y is (P) is the picture in color? (A) is the picture in color? (A) is the picture in color? (A) is the picture in color? A is y is (P) is the picture in color? (A) is the picture in color? (A) is the picture in color? (A) is the picture in color? A is y is (P) is the picture in co	1	Q: does the shelf have many layers?			
A: yes	1	A: just 2 that i can see			
A: yss Q: is the shelf full?		Q: is the picture in color?			
Pice is the shelf full? A: yes Or is the shelf full full full full full full full f		A: yes	O: is the shelf full?	A: yes	A: yes
A: yes ar burch of vegatables look fresh? A: yes ar burch of vegatables look fresh? A: yes ar burch of vegatables look fresh? A: yes ar burch arrows? A: yes ar burch arrows? A: yes ar burch arrows? A: yes ar burch arrows? A: yes are there carrows? A: yes are there prices? A: yes bis the any people? A: yes bis the any people? A: yes bis the price? A: yes bis the price? A: yes bis the bis bis bis bis the yes bis the yes bis the bis bis bis bis the yes bis the bis bis bis bis the yes bis the yes bis the bis		O: is the shelf full?			
Q: do the vegetables look fresh? A: yes Q: is the repetations look item? Q: is the repetations look item? a lumch of vegetables on a hier look of the list of vegetable has the most mumber present? Q: is the repetations look item? Q: is the repetations look item? Q: is the repetations look item? a lumch of vegetables on a hier of vegetable has the most number present? Q: what kind of vegetable has the most number present? Q:	and Carley Contraction		O: do the vegetables look fresh?	A: yes	A: yes
a hunch of vogetables on a A: yes Q: are there carrols? Q: are there prices? Q: are there any brick? Q: are there prices? Q: are there any brick? Q: are there	C C C C C C C C C C C C C C C C C C C				
9. are there carbod? A: no mostly seguides 9. are there carbody? 9. are there carbody? a bunch of vegetables on a shelf at the market. 9. are three carbody?? 9. are three carbody?? 9. are three carbody?? a bunch of vegetables on a shelf at the market. 9. are three carbody?? 9. are three carbody?? 9. are three carbody?? a bunch of vegetable has the most number present? 9. are three prices? 9. are three prices? 9. are three prices? A is soft at the market. 9. are three prices? 9. are three prices? 9. are three prices? 9. are three prices? A is soft at the market. 9. are three prices? 9. is the photo in color? 9. is the photo in color? 9. is the photo in color? 9. is three any people? 9. is	Et States	A von	Or any those connects?	A: yes	A: yes
a bunch of vegetables on shelf at the market Å: not that is see a bunch of vegetable has the nost number most number present? Å: no bunch of vegetable has the most number present?				Q: are there carrots?	Q: are there carrots?
a bunch of vegetables on shelf at the market A. Ino that is def market at the market P. are there cabbages? A in the cabbages? A ino or what kind of vegetable has the most number present? A sugar snap beams Grave there labels on the shell? A sugar snap beams Grave there shell Grave there there shell? A sugar snap beams Grave there snap beams Gra			A: no mostly veggles	A: yes	A: yes
a bunch of vegetable on a shelf at the market At no what kind of vegetable has the most number present? At no what has no what no wh					O: are there cabbages?
a. bird files accoss a lake next Q: what kind of vegetable has the most number present? Q: what kind of vegetable has the most number present? Q: what kind of vegetable has the most number present? A: sugar snap beams Q: are there labels on the shell? A: sys Q: are there labels on the shell? A: sys Q: are there prices? A: sys Q: is the photo in color? A: yes Q: is the photo in color? A: yes Q: is the photo in color? A: yes Q: is the photo in color? A: yes Q: is the photo in color? A: yes Q: is the photo in color? A: yes Q: is the photo in color? A: yes Q: is the photo in color? A: yes Q: is the photo in color? A: yes Q: is the photo in color? A: yes Q: is the photo in color? A: yes Q: is the photo in color? A: yes Q: is the photo in color? A: yes Q: is the photo in color? A: yes Q: is the photo in color? A: yes Q: is the photo in color? A: yes Q: is the photo in color? A: yes Q: is the photo in color? A: yes Q: is in tight? Q: is in tight? A: in o D: is in tight? Q: is in tight?					
a bird filter across a lake next 0; is the photo in color? 0; is the photo in color? <td< td=""><td>shelf at the market</td><td></td><td></td><td>O: what kind of vegetable has the</td><td>O: what kind of vegetable has the</td></td<>	shelf at the market			O: what kind of vegetable has the	O: what kind of vegetable has the
A: Squar, hubbane A:				most number present?	most number present?
A: sugar sup beams Q: are there labels on the shell? Q: are there prices? A: yes Q: are there prices? A: yes Q: are there prices? A: yes A: yes Q: are there prices? A: yes Q: are there prices? A: yes A: yes Q: are there prices? A: yes Q: are there prices? A: yes A: yes Q: is the photo in color? A: yes A: yes A: yes A: is there any people? Q: is there any people? A: is there any people? A: is there any people? A: is there any people? A: is there any people? A: is there any people? A: is there any people? A: is there any people? A: is the photo in color? A: is the own may binds? A: is there any people? A: is there any people? A: is the photo in color? A: is the own may binds? A: is the own may binds? A: is the own? A: is the own? A: is the own? A: is the own? A: is the own? A: is the own? A: is the own? A: is the own? A: is the own? A: is the own? A: is the own? A: is the own? A: is the own? A:			A: squash, potatoes	At i don't know	At sugar span beans
Q are there indee: on the shell? A 'yes Q are there prices? Q is the photo in color? Q is there any people? Q is the photo in color? Q is the		A: sugar snap beans	Q: are there labels on the shelf?		
A: yes Q: are there prices? A: yes Q: are there prices? A: yes Image: the problem is accord? A: yes Q: is the photo in color? A: yes Image: the problem is the color is the color? A: no Q: is the photo in color? A: yes Q: is the photo in color? A: yes Image: the photo in color? A: no A: no A: no A: no A: no Image: the photo in color? A: no A: no A: no A: no A: no Image: the photo in color? A: no A: no A: no A: no A: no Image: the photo in color? A: no A: no A: no A: no A: no Image: the photo in color? A: no A: no A: no A: no A: no Image: the photo in color? A: no A: no A: no A: no A: no Image: the photo in color? A: no A: no A: no A: no A: no Image: the photo in color? A: yes A: yes A: yes A: yes A: yes <td< td=""><td></td><td>O: are there labels on the shelf?</td><td>A: yes</td><td></td><td></td></td<>		O: are there labels on the shelf?	A: yes		
Qi are there prices? Ai no Vi are intere prices? Vi are intere prices? Vi are intere prices? Vi are intere prices? Vi yes Qi is the photo in color? Qi is t		A: ves	O: are there prices?		
A: yes A: yes A: yes A: yes A: yes Q: is the photo in color? Q: is there any people? Q: is the photo in color? Q: is there any people? Q: is the photo in color? Q: are there many buildings? Q: are there many buildings? Q: are there many buildings? Q:					
Q: is the photo in color? A: yes Q: is the photo in color? A: yes Q: is the photo in color? A: yes A: yes Q: is the photo in color? A: yes Q: is the photo in color? A: yes A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no is the bird brown? A: in diff. A: no A: no A: no is the bird brown? A: in diff. A: no A: no A: no A: no is the bird brown? A: in diff. A: no A: no A: no A: no is the dire brown? A: in diff. A: no A: no A: no A: no is the dire brown? A: in diff. A: no A: no A: no A: no A: no is the dire brown? A: no A: no A: no A: no A: no A: no A: no is the dire brown? A: no		O: are there prices?	A: no		
Ai yes Ai yes Ai yes Ai yes Ai yes Ai yes a bird flips across a lake next is is there any popole? Qi is there any popole? Qi is there any popole? Qi is there any popole? a bird flips across a lake next is in glip1? Qi is the any popole? Qi is there any popole? Qi is there any popole? a bird flips across a lake next is in glip1? Qi is the glip1? Qi is the glip1? Qi is the glip1? a bird flips across a lake next is it day? Qi is the glip1? Qi is the glip1? Qi is the glip1? a bird flips across a lake next is the sint flip? Qi is the glip1? Qi is the glip1? Qi is the glip1? Ai yes Ai yes Ai yes Ai yes Ai yes Ai yes a bird flips across a lake next is the sint yes Ai yes Ai yes Ai yes Ai yes Ai yes Ai yes Ai yes Ai yes Ai yes a lot of them Qi are there trees? Ai yes Ai yes Ai yes Ai yes Ai yes Ai yes Ai yes Ai yes Ai yes Qi are there trees? Ai yes Ai yes Ai yes			A: no		
Provide and the second of the secon		A: yes		À: yes	Ă: yes
Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no a bird flies across a lake next is it day? Qi is the bird hown? Qi is it day? Qi is it day? Air no Air no Air no a bird flies across a lake next is it day? Qi is the sky visible? Qi are there trees? Qi		A: yes		À: yes	Ă: yes
Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no Air no a bird flies across a lake next is it day? Qi is the bird hown? Qi is it day? Qi is it day? Air no Air no Air no a bird flies across a lake next is it day? Qi is the sky visible? Qi are there trees? Qi		A: yes Q: is the photo in color?	Q: is the photo in color?	A: yes Q: is the photo in color?	A: yes Q: is the photo in color?
Q: how many binds? Q: how how many binds? Q: how how many binds? Q: how how? Q: how how many binds? Q: how how how? Q: how		A: yes Q: is the photo in color? A: yes	Q: is the photo in color? A: yes	A: yes Q: is the photo in color? A: yes	A: yes Q: is the photo in color? A: yes
A: 1 A: 1 A: 1 A: 1 A: 1 a bird flies across a lake next to some buildings A: is in dight? A: is in dight? A: is in dight? A: is in dight? a bird flies across a lake next to some buildings A: no A: no A: no A: no A: no a bird flies across a lake next to some buildings A: no A: no A: no A: no A: no a bird flies across a lake next to some buildings A: no A: no A: no A: no A: no a bird flies across a lake next to some buildings A: no A: no A: no A: no A: no a inf dlies across a lake next to some buildings A: no A: no A: no A: no A: no a inf dlies across a lake next to some buildings A: no A: no A: no A: no A: no a inf dlies across a lake next to some buildings A: no A: no A: no A: no A: no a rule nerve nany buildings? A: no A: no A: no A: no A: no a in of nerve nany buildings? A: no A: no A: no A: no A: no A: no A: no		A: yes Q: is the photo in color? A: yes Q: is there any people?	Q: is the photo in color? A: yes Q: is there any people?	A: yes Q: is the photo in color? A: yes Q: is there any people?	A: yes Q: is the photo in color? A: yes Q: is there any people?
Q: is it night? Q: is it hight? Q: is it hight? Q: is it hight? Q: is it night? Q: is it night? Q: is it night? Q: is it hight? Q: is the bird brown? Q: is the bird?		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no	Q: is the photo in color? A: yes Q: is there any people? A: no	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no
A: no A: no <td< td=""><td></td><td>A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds?</td><td>Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds?</td><td>A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds?</td><td>A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds?</td></td<>		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds?	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds?	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds?	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds?
4 Q: is the bird brown? Q: is the bird? Q: are there trees? Q: are there many buildings? Q: are there many buildings? Q: are three trees? Q: are three trees? Q: are there trees?		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1
A in oit while A in on A in one A in one a bird flies across a lake next A in oit while A in off A is it day? A is it day? a bird flies across a lake next A is yes A iyes A iyes A iyes a bird flies across a lake next A is yes A iyes A iyes A iyes a bird flies across a lake next A is the threes? A is no if the day visible? A is the day visible? A is no if them A is yes A iyes A is the day visible? A is the day is the day is the day visible? A is the day is the day visible? A is the day is the day is the day visible? A is the day is the		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is it night?	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in right?	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is it night?	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is it night?
a bird flies across a lake next to some buildings Q: is it day? Q: is the day is it day? Q: are there trees? Q: are there trees? Q: are there trees? Q: are there trees? Q: are there many buildings? A: ros A: ros <td< td=""><td></td><td>A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in right? A: no</td><td>Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is night? A: no</td><td>A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in third A: no it is daytime</td><td>A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in right? A: no</td></td<>		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in right? A: no	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is night? A: no	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in third A: no it is daytime	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in right? A: no
Air yes Air yes Air yes Air yes a bird flies across a lake next Qi are there trees? Air yes Air yes a bird flies across a lake next Qi are there trees? Air yes Air yes a bird flies across a lake next Qi are there trees? Air yes Air yes a bird flies across a lake next Qi are there trees? Air yes Air yes a bird flies across a lake next Qi are there many buildings? Air no Air yes Air no Air no Air yes Air yes Air yes are there many buildings? Air no Air yes Air yes Air yes Air a lot of them Qi are there many buildings? Air yes Air yes Air yes Air yes Air yes Air yes Air yes Air yes Air yes Air yes Air yes Air yes Air yes Air yes Air yes Air yes Air yes Air yes Air yes Air yes Air yes Qi are there slush? Qi are there nany buildings? Air yes Air yes Air yes Qi are there slush? Qi are there slush? Air y		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is night? A: no Q: is the inght? A: no Q: is the first brown?	Q: is the photo in color? A: yes Q: is there any people? A: no Q: is in night? A: no Q: is the init brown?	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 10 Q: is the jint from the photom of the ph	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is the first first? A: no Q: is the first brown?
a bird files across a lake next, to some buildings Q: are there trees? Q: are there many buildings? A: yes A:		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in right? A: no Q: is the bird brow?? A: no if's white	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in right? A: no Q: is the bird brown? A: no	A: yes A: yes A: yes A: no A: no	A: yes (); is the photo in color? (); yes (); is there any people? (); is there any people? (); is now many birds? (); is no photo model (); is nophoto m
a bird flying over the ocean? a bird flying over the ocean? a bird flying over the ocean? a bird flying over the ocean? beach at sunrise beach beach at sunrise beach at sunrise beach beach at sunrise beach beach at sunrise beach beach at sunrise beach beach at sunrise beach at sunrise beach beach at sunrise beach beach at sunrise beach beach at sunrise beach beach at sunrise beach at sunri		A: yes Q: is the photo in color? A: yes Q: is there any people? A: 10 Q: how many birds? A: 10 Q: is in right? A: no Q: is the bird brown? A: no if's white Q: is the bird brown? A: no if's white Q: is the day?	Q: is the photo in color? A: yes Q: is there any people? A: to Q: how many birds? A: 1 Q: is in right? A: no Q: is it bird brown? A: no Q: is day?	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 10 Q: is whith? A: no it is daytime Q: is the bird brow? A: nope Q: is the day?	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is thight? A: no Q: is in tight? A: no Q: is the distribution on? A: no Q: is the distribution on? A: no Q: is the distribution on? A: no Q: is the distribution of the dist
to some buildings Or is the sky visible? Or is the read the sky visible? Or is the read the sky visible? Or is the sky visible?		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in night? A: no Q: is the bird brow? A: no if's white Q: is it day? A: yes	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in right? A: no Q: is the bird brown? A: no Q: is it day? A: yes	A: yes () is the photo in color? () iyes ()	A: yes Q: is the photo in color? Q: is yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is tright? A: no Q: is the bird brown? A: no Q: is it day? A: yes
A no		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: no Q: is in right? A: no Q: is in right? A: no Q: is the bird brown? A: no if s white Q: is it day? A: yes Q: arc there trees?	Q: is the photo in color? A: yes Q: is there any people? A: to Q: is there any people? A: to Q: is in right? A: no Q: is the bird brown? A: no Q: is it aday? A: yes Q: are there trees?	A: yes Q: is the photo in color? A: yes O: is there any people? A: no O: how many birds? A: no O: is night? A: no it is daytime O: is the bird brow? A: nope O: is the dist brow? A: nope O: is the dist brow? A: yes O: is the dist br	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 10 Q: how many birds? A: no Q: is it night? A: no Q: is the bird brown? A: no Q: is the bird brown? A: no Q: is the dirdy? A: yes Q: are there trees?
Qr are there many buildings?		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 10 Q: is the phot birds? A: no Q: is the bird brow?? A: no if's white Q: is it day? A: yes Q: are there trees? A: yes	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in right? A: no Q: is the bird brown? A: no Q: is it day? A: yes Q: are there trees? A: yes	A: yes Q: is the photo in color? V: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in night? A: no It is day(ime Q: is it day? A: yes Q: an there there's? A: yes	A: yes Q: is the photo in color? Q: is the photo in color? Q: is the any people? A: no Q: how many birds? A: 1 Q: is the joint box many birds? A: 1 Q: is the bird brown? A: no Q: is it day? A: yes Q: are there trees? A: yes
Ar a lot of them Ar yes Ar yes Ar yes Ar yes Ar yes Qr are the trees lush? Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Qr are the trees lush? Ar yes Qr are the trees lush? Ar yes Yes Yes Qr any ose clouds? Qr is the photo in color? Qr is the photo in sclor? Yes Yes Yes Qr any ose clouds? Qr is the photo in scloor? Qr what color is the ocean? Yes Yes Yes Qr what color is the ocean? Qr what color is the ocean? Yes Yes Yes Yes Qr any you see the sand on the Yes Yes Yes Yes Yes Yes Qr any people around? Yes Yes Yes Yes Yes Yes Ar yes Yes Yes Yes Yes Yes Yes Yes Qr any yeople around? Yes Yes <t< td=""><td></td><td>A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: is wow many binds? A: no first white? Q: is in finght? Q: is in finght? Q: is the bind brown? A: no firs white? Q: is in day? Q: are there trees? A: yes Q: is the ky visible?</td><td>Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is it night? A: no Q: is it night? A: yes Q: are there trees? A: yes Q: is the key visible?</td><td>A: yes Q: is the photo in color? A: yes Q: is there any people? A: no P: how many birds? A: n Q: is with a straight? A: no it is daytime Q: is the bird brow? A: nope Q: is the bird brow? A: nope Q: arc here trees? A: yes Q: arc here trees? A: yes Q: is the straight? A: yes Q: arc here trees? A: yes Q: is the straight? A: yes Q: arc here trees? A: yes Q: is the straight? A: yes Q: is the strai</td><td>A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 10 Q: is night? A: no Q: is the bird brown? A: no Q: is the bird brown? A: no Q: is the bird brown? A: yes Q: is the tree trees? A: yes Q: is the ky visible?</td></t<>		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: is wow many binds? A: no first white? Q: is in finght? Q: is in finght? Q: is the bind brown? A: no firs white? Q: is in day? Q: are there trees? A: yes Q: is the ky visible?	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is it night? A: no Q: is it night? A: yes Q: are there trees? A: yes Q: is the key visible?	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no P: how many birds? A: n Q: is with a straight? A: no it is daytime Q: is the bird brow? A: nope Q: is the bird brow? A: nope Q: arc here trees? A: yes Q: arc here trees? A: yes Q: is the straight? A: yes Q: arc here trees? A: yes Q: is the straight? A: yes Q: arc here trees? A: yes Q: is the straight? A: yes Q: is the strai	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 10 Q: is night? A: no Q: is the bird brown? A: no Q: is the bird brown? A: no Q: is the bird brown? A: yes Q: is the tree trees? A: yes Q: is the ky visible?
Ar a lot of them Ar yes Ar yes Ar yes Ar yes Ar yes Qr are the trees lush? Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Ar yes Qr are the trees lush? Ar yes Qr are the trees lush? Ar yes Yes Yes Qr any ose clouds? Qr is the photo in color? Qr is the photo in sclor? Yes Yes Yes Qr any ose clouds? Qr is the photo in scloor? Qr what color is the ocean? Yes Yes Yes Qr what color is the ocean? Qr what color is the ocean? Yes Yes Yes Yes Qr any you see the sand on the Yes Yes Yes Yes Yes Yes Qr any people around? Yes Yes Yes Yes Yes Yes Ar yes Yes Yes Yes Yes Yes Yes Yes Qr any yeople around? Yes Yes <t< td=""><td></td><td>A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: no Q: how many birds? A: no Q: is the bird brow?? A: no if's white Q: is it day? A: yes Q: are there trees? A: yes Q: is the sky visible? A: no</td><td>Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in right? A: no Q: is the bird brown? A: no Q: is the bird brown? A: no Q: is tay? A: yes Q: as the sky visible? A: no</td><td>A: yes Qr: is the photo in color? V: yes Qr: is there any people? A: no Qr: how many birds? A: 1 Qr: is night? A: no ric advyime Qr: is the bird brown? A: nope Qr: is it day? A: yes Qr: and there show res? A: yes Qr: is the sky visible? A: yes</td><td>A: yes Q: is the photo in color? V: yes A: no Q: how many birds? A: 1 Q: is the jan bow many birds? A: 1 Q: is the bird brown? A: no Q: is the bird brown? A: no Q: is it day? A: yes Q: as the sky visible? A: yes</td></t<>		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: no Q: how many birds? A: no Q: is the bird brow?? A: no if's white Q: is it day? A: yes Q: are there trees? A: yes Q: is the sky visible? A: no	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in right? A: no Q: is the bird brown? A: no Q: is the bird brown? A: no Q: is tay? A: yes Q: as the sky visible? A: no	A: yes Qr: is the photo in color? V: yes Qr: is there any people? A: no Qr: how many birds? A: 1 Qr: is night? A: no ric advyime Qr: is the bird brown? A: nope Qr: is it day? A: yes Qr: and there show res? A: yes Qr: is the sky visible? A: yes	A: yes Q: is the photo in color? V: yes A: no Q: how many birds? A: 1 Q: is the jan bow many birds? A: 1 Q: is the bird brown? A: no Q: is the bird brown? A: no Q: is it day? A: yes Q: as the sky visible? A: yes
Q: are the trees lush? Q: are the trees lush?<		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in inght? A: is in inght? A: is inght? A: yes Q: is in the white Q: is the key visible? A: no Q: are there many buildings?	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in inght? A: no Q: is in tagy? Q: are there trees? Q: yes Q: so they visible? A: no Q: are there trees? Q: so they visible? A: no Q: are there many buildings?	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in ight? A: no Q: is in kind throw? Q: is the bird throw? A: nope Q: is the bird throw? A: yes Q: are there threes? A: yes Q: set here three? A: yes Q: are there threes? A: yes Q: are there threes? A: yes Q: are there threes? A: yes Q: are there many buildings?	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: no Q: is night? A: no Q: is the bird brown? A: yes Q: are there trees? A: yes Q: is the sky visible? A: yes Q: are there trees? A: yes A: yes
Air yes Air yes Air yes Air yes Air yes Q: is the photo in color? Q: can you see clouds? Q: what color is the ocean? Q: what color is the sand on the beach? Deach?		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: no Q: how many birds? A: no Q: is the bird brow? A: no if's white Q: is it day? A: yes Q: are there trees? A: yes Q: is the sky visible? A: no Q: is the sky visible? A: no A: no there there many buildings? A: a lot of them	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in right? A: no Q: is the bird brown? A: no Q: are there trees? A: yes Q: are there many buildings? A: no Q: are there many buildings?	A: yes Q: is the photo in color? A: yes Q: is there any people? A: is on the show many birds? A: no ti is daytime Q: is the bird brown? A: nope Q: is the bird brown? A: ange Q: is the bird brown? A: yes Q: ar there it day? A: yes Q: ar there trees? A: yes Q: ar there would be any constraint of the show is blow? A: yes Q: are there many buildings? A: yes	A: yes Q: is the photo in color? A: yes A: is there any people? A: no A: no A: is in right? A: no Q: is the bird brown? A: no Q: is the bird brown? A: no Q: is it day? A: yes Q: ar there there trees? A: yes Q: ar there any buildings? A: yes Q: ar there many buildings?
Q: is the photo in color? A: yes is the color is the ocean? A: is brownist gray A: is brownist gray A: is brownist gray A: in or use the sand on the beach? A: in or use the sand on the beach? A: in or is the bird? A: in or is the bird? A: white beach at sumrise beach at sumrise beach at sumrise beach at sumrise Q: is the photo in color? A: is photo set the sand on the beach? A: is photo set the sand on the beach? A: is photo color is the bird? A: white A: white A: white A: white A: white A: white A: is on A: is o		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: no Q: how many birds? A: no Q: is the bird brow? A: no if's white Q: is it day? A: yes Q: are there trees? A: yes Q: is the sky visible? A: no Q: is the sky visible? A: no A: no there there many buildings? A: a lot of them	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in right? A: no Q: is the bird brown? A: no Q: are there trees? A: yes Q: are there many buildings? A: no Q: are there many buildings?	A: yes Q: is the photo in color? A: yes Q: is there any people? A: is on the show many birds? A: no ti is daytime Q: is the bird brown? A: nope Q: is the bird brown? A: ange Q: is the bird brown? A: yes Q: ar there it day? A: yes Q: ar there trees? A: yes Q: ar there would be any constraint of the show is blow? A: yes Q: are there many buildings? A: yes	A: yes Q: is the photo in color? A: yes A: is there any people? A: no A: no A: is in right? A: no Q: is the bird brown? A: no Q: is the bird brown? A: no Q: is it day? A: yes Q: ar there there trees? A: yes Q: ar there any buildings? A: yes Q: ar there many buildings?
A: yes A: yes A: yes A: yes A: yes A: yes C: an you see clouds? A: yes A: yes A: yes A: yes C: an you see clouds? A: yes A: yes A: yes A: yes C: an you see clouds? A: yes A: yes A: yes A: yes C: an you see clouds? A: yes A: yes A: yes A: yes C: an you see clouds? A: yes A: yes A: yes A: grey What color is the ocean? A: what color is the ocean? A: yes A: grey C: an you see the sand on the beach? A: in is brownish gray A: yes B: and flying over the ocean? A: yes A: yes A: yes Q: any poeple around? A: no A: no A: no A: no A: no A: no A: no A: no Beach at sumrise A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no A: no </td <td></td> <td>A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is the ingln? A: is ingln? A: yes Q: is it duy? A: yes Q: are there trees? A: yes Q: are there many buildings? A: at of them Q: are there sulsh?</td> <td>Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in inght? A: no Q: is in the bird brown? A: no Q: is in tady? Q: is in tady? Q: are there trees? Q: is the dxy vasible? A: no Q: are there many buildings? A: yes Q: are there trees lash?</td> <td>A: yes Q: is the photo in color? A: yes Q: is there any people? A: no C: how many birds? A: n Q: is with a there any people? A: no C: is the light from? A: nope Q: is the bird how? A: yes Q: are there trees? A: yes Q: are there trees light? A: yes Q: are there light? A: yes Q: are t</td> <td>A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: n Q: is n' night? A: no Q: is it night? A: no Q: is the bird brown? A: no Q: is the bird brown? A: yes Q: arc here trees? A: yes Q: arc here trees hyper hyper</td>		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is the ingln? A: is ingln? A: yes Q: is it duy? A: yes Q: are there trees? A: yes Q: are there many buildings? A: at of them Q: are there sulsh?	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in inght? A: no Q: is in the bird brown? A: no Q: is in tady? Q: is in tady? Q: are there trees? Q: is the dxy vasible? A: no Q: are there many buildings? A: yes Q: are there trees lash?	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no C: how many birds? A: n Q: is with a there any people? A: no C: is the light from? A: nope Q: is the bird how? A: yes Q: are there trees? A: yes Q: are there trees light? A: yes Q: are there light? A: yes Q: are t	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: n Q: is n' night? A: no Q: is it night? A: no Q: is the bird brown? A: no Q: is the bird brown? A: yes Q: arc here trees? A: yes Q: arc here trees hyper
A: yes A: yes A: yes A: yes A: yes A: yes A: yes Q: can you see clouds? Q: can you see clouds? Q: can you see clouds? Com you see clouds? A: yes A: yes A: yes A: yes A: yes A: yes Wind color is the ocean? Q: can you see clouds? Q: can you see clouds? Q: can you see clouds? Q: what color is the ocean? Q: can you see the sand on the beach? Q: can you see the sand on the beach? Q: can you see the sand on the beach? Q: can you see the sand on the beach? A: yes A: yes A: yes A: yes A: on A: yes A: yes A: on A: on A: yes A: on A: on A: on A: white A: white A: white A: white A: white A: white A: white A: white A: white A: white A: on A: on A: on A: on A: on A: on A: on A: on A: on A: on A: on you see any boats? A: on A: on A: on A: on A: on A: on A: on A: on A: on A		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: no Q: is the sind birds? A: no Q: is the birds brown? A: no if's white Q: is it day? A: yes Q: are there trees? A: job of them Q: is no they visible? A: a lot of them Q: are the trees lush? A: yes	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in inght? A: no Q: is in the bird brown? A: no Q: is in tady? Q: is in tady? Q: are there trees? Q: is the dxy vasible? A: no Q: are there many buildings? A: yes Q: are there trees lash?	A: yes (): is the photo in color? A: yes Q: is three any people? A: no Q: how many birds? A: no Q: how many birds? A: no it is day? A: no it is day? A: nope Q: is it day? A: yes Q: are three stres? A: yes Q: is the sky visible? A: yes Q: are three threes lush? A: are the trees lush? A: are the tree	A: yes (): is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: no Q: how many birds? A: no Q: is the bird brown? A: no Q: is the bird brown? A: yes Q: as there are there trees? A: yes Q: is the sky visible? A: yes Q: are there many buildings? A: yes Q: are there trees lush? A: yes
Q: can you see clouds? Q: what color is the ocean? Q: what color is the sand on the beach? Deach? Deach? Q: can you see the sand on the beach? Deach? Deac		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: is the bird brown? A: no it's white Q: is the bird brown? A: no it's white Q: is the bird brown? A: no it's white Q: is the bird brown? A: yes Q: as the bird brown? A: yes Q: as the bird brown? A: yes Q: as the bird brown? A: yes Q: are there many buildings? A: yes Q: are there many buildings? A: yes Q: are the trees lush? A: yes	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in inght? A: no Q: is in tagy? Q: are there trees? Q: is no Q: is the day visible? A: no Q: are there trees? Q: are there many buildings? A: yes Q: are there slush? A: yes	A: yes Q: is the photo in color? A: yes A: yes A: here any people? A: how many birds? A: how many birds? A: how fin dytime Q: is the bird brown? A: nop: A: nop: Q: is the bird brown? A: nop: Q: is the bird brown? A: yes Q: ar there trees? A: yes Q: ar there theres? A: yes Q: ar there slush? A: yes	A: yes Q: is the photo in color? A: yes A: yes A: is there any people? A: is there any people? A: is the moment A: is in right? A: is the bird brown? A: is the bird brown? A: is it day? A: yes Q: is the sky visible? A: yes Q: is the sky visible? A: yes Q: are there nearly buildings? A: yes Q: is the photo in color?
A: yes A: yes A: yes A: yes A: yes A: yes What color is the occan? Q: what color is the occan? A: grey What color is the occan? Q: what color is the occan? Q: what color is the occan? Q: what color is the occan? A: grey Q: can you set the sand on the beach? Q: any you set the sand on the beach? Q: any you set the sand on the beach? Q: any you set the sand on the beach? beach? Q: any people around? A: yes A: yes A: yes A: yes Q: any people around? A: yes A: yes A: yes A: yes A: yes a bird flying over the occan? Q: what color is the bird? A: no A: no A: no A: no a bird flying over the occan? Q: what color is the bird? A: no A: no A: no A: no Q: any people around? A: no Beach a: summise Q: any other birds? Q: any other birds? Q: any other birds? A: no		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is the ind prown? A: no if's white Q: is the bird brown? A: yes Q: as there rese? A: yes Q: as the first wishle? A: so if's hort fhem Q: are there many buildings? A: yes Q: as the first bury bird? A: so if the photo in color? A: yes	Q: is the photo in color? A: yes Q: is there any people? A: no Q: now many birds? A: 1 Q: is tinght? A: no Q: is the bird brown? A: no Q: is the bird brown? A: no Q: is the bird brown? A: ros Q: is the bird brown? A: yes Q: are there reses? A: yes Q: are there many buildings? A: yes Q: are the trees lush? A: yes Q: is the photo in color?	A: yes (): is the photo in color? A: yes (): is there any people? A: no (): how many birds? A: no (): how many birds? A: no (): is duy? (): is the bird brown? A: nopti (): is the bird brown? A: nopti (): is the bird brown? A: spectra (): is the bird brown? A: yes (): is the sky visible? A: yes (): is the sky visible? A: yes (): is the photo in color? A: yes (): is the photo in color? A: yes (): is the photo i	A: yes () () () () () () () () () () () () ()
Q: what color is the occan? Q: can you see the sand on the beach? Q: can you see the sand on the beach? D: what color is the point? Q: what color is the occan? Q: what color is the oc		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: is there any people? A: no Q: is there any people? A: no Q: is the hird brown? A: no it's white Q: is the bird brown? A: no it's white Q: is the sky visible? A: or there may buildings? A: a lot of them Q: are there slush? A: yes Q: is the photo in color? A: yes Q: are the tores lush?	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in inght? A: no Q: is in the bird brown? A: is inght? Q: is in tady? Q: is the ky visible? A: yes Q: are there many buildings? A: yes Q: are theres lash? A: yes	A: yes Q: is the photo in color? A: yes C: sthere any people? A: of the photo in color? A: of is day: A: no fis day: A:	A: yes Q: is the photo in color? A: yes C: is there any people? A: is C: is uncertainty people? A: is A: is
Ai grey Qi winit coord is the doctain? Ai ti is brownish gray Ai grey Ai grey Can you see the sand on the beach? Qi can you see the sand on the beach?<		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: is there any people? A: no Q: is there any people? A: no Q: is the hird brown? A: no it's white Q: is the bird brown? A: no it's white Q: is the sky visible? A: or there may buildings? A: a lot of them Q: are there slush? A: yes Q: is the photo in color? A: yes Q: are the tores lush?	Q: is the photo in color? A: yes Q: is there any people? A: no Q: now many birds? A: 1 Q: is three any people? A: no Q: is three any people? A: no Q: is the bird brown? A: no Q: is the bird brown? A: yes Q: ar there rany buildings? A: yes Q: ar there many buildings? A: yes Q: is the photo in color? A: yes Q: ary to use clouds?	A: yes Q: is the photo in color? A: yes C: sthere any people? A: of the photo in color? A: of is day: A: no fis day: A:	A: yes Q: is the photo in color? A: yes C: is there any people? A: is C: is uncertainty people? A: is A: is
Q: can you see the sand on the beach? Deach? Q: can you see the sand on the beach? Deach? Q: can you see the sand on the beach? Deach? Q: can you see the sand on the beach? Deach? Q: can you see any beats? Q: do you see		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is the bird brown? A: no if's white Q: is the bird brown? A: yes Q: as the bird brown? A: yes Q: as there rese? A: yes Q: as there many buildings? A: yes Q: as the freemany buildings? A: yes Q: as the press lush? A: yes Q: as the stor samy buildings?	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is the bird brown? A: no Q: is the sky visible? Q: is the sky visible? Q: are there many buildings? Q: are there shush? A: yes Q: is the photo in color? A: yes Q: concours any use clouds? A: yes	A: yes (): is the photo in color? A: yes (): is there any people? A: no (): how many birds? A: no (): how many birds? A: no (): is duy? A: no (): is duy? A: no (): is duy? A: no (): is duy? A: yes (): is the bird brown? A: no (): is the bird brown? A: yes (): is the sky visible? A: yes (): is the photo in color? A: yes (): is (): on you see clouds? A: yes (): is (): on the sky visible? (): on the reses lush? (): on the reses lush? (): on you see clouds? (): yes (): on you see clouds? (): on you see c	A: yes () () () () () () () () () () () () ()
beach? beach? beach? beach? a bird flying over the ocean beach at sunrise A: yes A: yes A: yes a bird flying over the ocean beach at sunrise Q: any people around? Q: any people around? Q: any people around? a bird flying over the ocean beach at sunrise Q: doy ou see any boats? Q: what color is the bird? Q: what color is the bird? Q: what color is the bird? Q: doy ou see any boats? A: no A: no A: white A: white a bird flying over the ocean beach at sunrise Q: doy ou see any boats? A: white A: white Q: any other birds? A: no A: no A: no A: no A: no Q: any other birds? A: no A: no A: no A: in o Q: any other birds? A: no A: no A: no A: 1 other Q: any other birds? A: no A: no A: no A: 1 other Q: any igts or balloons? A: no A: no A: no A: no Q: any igts or balloons? A: no A: no A: no Q: do you see buildings? Q: do you se		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: is the bird brown? A: no it's white Q: is the bird brown? A: no it's white Q: is the bird brown? A: yes Q: are there res? A: yes Q: are there many buildings? A: a lot of them Q: are the trees lush? A: yes Q: are the too in color? A: yes Q: is the photo in color? A: yes Q: ary ou see clouds? A: yes Q: can you see clouds? A: yes Q: what color is the ocean?	Q: is the photo in color? A: yes Q: is there any people? A: no Q: is there any people? A: no Q: is there any people? A: no Q: is three any people? A: no Q: is the bird brown? A: no Q: is it day? A: yes Q: as the sky visible? A: no Q: are there renes? A: yes Q: are there rany buildings? A: yes Q: is the photo in color? A: yes Q: can you see clouds? A: yes Q: can you see clouds? A: yes Q: can you see clouds? A: yes Q: con you see clouds? A: yes <td>A: yes Q: is the photo in color? A: yes Q: is there any people? A: how many birds? A: how many birds? A: is in ight? A: is in dist. A: no fi is dytime Q: is the bird brown? A: no fi is dytime Q: is the bird brown? A: is is iday? Q: is the sky visible? A: yes Q: are there many buildings? A: yes Q: are there slush? Q: is the photo in color? A: yes Q: can you see clouds? A: yes Q: what color is the ocean?</td> <td>A: yes Q: is the photo in color? A: yes C: shere any people? A: how many birds? A: how many birds? A: how many birds? A: is the bird brown? A: is the bird brown? A: is it aly? A: is the bird brown? A: is it aly? A: is the bird brown? A: is it aly? A: is the bird brown? A: is the sky visible? A: yes Q: are there many buildings? A: yes Q: is the photo in color? A: yes Q: can you see clouds? A: yes Q: what color is the ocean?</td>	A: yes Q: is the photo in color? A: yes Q: is there any people? A: how many birds? A: how many birds? A: is in ight? A: is in dist. A: no fi is dytime Q: is the bird brown? A: no fi is dytime Q: is the bird brown? A: is is iday? Q: is the sky visible? A: yes Q: are there many buildings? A: yes Q: are there slush? Q: is the photo in color? A: yes Q: can you see clouds? A: yes Q: what color is the ocean?	A: yes Q: is the photo in color? A: yes C: shere any people? A: how many birds? A: how many birds? A: how many birds? A: is the bird brown? A: is the bird brown? A: is it aly? A: is the bird brown? A: is it aly? A: is the bird brown? A: is it aly? A: is the bird brown? A: is the sky visible? A: yes Q: are there many buildings? A: yes Q: is the photo in color? A: yes Q: can you see clouds? A: yes Q: what color is the ocean?
A: yes A: yes A: yes A: yes a bird flying over the ocean beach at sunrise A: yes A: yes A: yes A: no a bird flying over the ocean A: no Q: any people around? Q: any people around? Q: any people around? a bird flying over the ocean beach at sunrise Vanta color is the bird? Q: what color is the bird? Q: what color is the bird? Q: what color is the bird? a bird flying over the ocean beach at sunrise Q: do you see any boats? A: no Q: any other birds? Q: any other birds? Q: any other birds? Q: any other birds? A: 1 other Q: any isto or balloons? Q: any other birds? Q: any isto r balloons? Q: any isto r balloons? Q: do you see buildings?		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is the part of the photo in color? A: no if's white Q: is the bird brown? A: no if's white Q: is the bird brown? A: yes Q: as the sky visible? A: as lot of there slush? A: yes Q: as the photo in color? A: yes Q: any use sec clouds? Y: yes Q: any the color is the ocean?	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is the bird brown? A: no Q: is there trees? A: yes Q: are there many buildings? Q: are there many buildings? A: yes Q: are there nonyous see clouds? A: yes Q: with color is the ocean? A: bue	A: yes () () () () () () () () () () () () ()	A: yes () () () () () () () () () () () () ()
Q: any people around?		A: yes Q: is the photo in color? A: yes Q: is three any people? A: no Q: is in right? Q: is in dray? Q: is in dray? Q: are three trees? Q: are three trees? Q: are three trees lush? A: set of them Q: is the photo in color? A: yes Q: can ye see clouds? A: yes decoins the ocean?	Q: is the photo in color? A: yes Q: is there any people? A: no Q: now many birds? A: 1 Q: is the photo in color? A: no Q: is the bird brown? A: no Q: is the bird brown? A: no Q: is it day? A: yes Q: are there trees? A: so Q: are there many buildings? A: yes Q: is the photo in color? A: yes Q: can you see clouds? A: yes Q: can you see clouds? A: yes Q: can you see the sand on the beach?	A: yes Q: is the photo in color? A: yes Q: is there any people? A: of is there any people? A: of its daytime Q: is or with a daytime Q: is the bird brown? A: is for its daytime Q: is the bird brown? A: of its daytime Q: is the bird brown? A: yes A:	A: yes Q: is the photo in color? A: yes Q: is there any people? A: is Q: is there any people? A: is is on the photo in color? A: is it night? A: is the bird brown? A: is A: is A: is the bird brown? A: is A: is
A: no A: no A: no A: no Q: what color is the bird? a bird flying over the ocean Q: do yous see any boats? A: white A: white beach at sunrise Q: do you see any boats? Q: do you see any boats? Q: do you see any boats? a bird flying over the ocean Q: any other birds? Q: any other birds? Q: any other birds? a: no Q: any other birds? A: no A: no A: 1 other A: no Q: any other birds? Q: any other birds? A: 1 other Q: any ifs or balloons? Q: any other birds? Q: any ifs or balloons? A: no Q: any ifs or balloons? Q: any ifs or balloons? Q: any ifs or balloons? A: no Q: do you see buildings? Q: do you see buildings? Q: do you see buildings?		A: yes Qr: is the photo in color? A: yes Qr: is there any people? A: no Qr: how many birds? A: 1 Qr: is the any people? A: no Qr: is the bird brown? A: no if's white Qr: is the bird brown? A: yes Qr: is the sky visible? A: no the sky visible? A: no the theres luah? A: yes Qr: is the photo in color? A: yes Qr: any nou see clouds? A: yes Qr: any nou see clouds? A: yes Qr: any nou see clouds? A: yes Qr: any nou see the sand on the beach?	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is the bird brown? A: no Q: is there rerees? A: yes Q: are there many buildings? Q: are there many buildings? A: yes Q: are there non you see clouds? A: yes Q: what color is the ocean? A: yes Q: what color is the ocean? A: bulk Q: en up ous see the sand on the beach? A: yes	A: yes () () () () () () () () () () () () ()	A: yes () () () () () () () () () () () () ()
Q: what color is the bird? Q: do you see any boats? Q: do you see any boats? Q: do you see any boats? Q: do you see any other birds? Q: do you see birdings? Q: d		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in the ingle? A: in ingle? Q: is in the ingle? A: no if's white Q: is in day? A: yes Q: are there trees? A: yes Q: are there many buildings? A: a lot of them Q: is the photo in color? Q: can be photo in sclor? Q: was calcouds? Q: was end by accounds? A: yes	Q: is the photo in color? A: yes Q: is there any people? A: no Q: now many birds? A: 1 Q: is there any people? A: no Q: now many birds? A: 1 Q: is in right? A: no Q: is it day? A: yes Q: are there renes? A: sos Q: are there many buildings? A: yes Q: is the photo in color? A: yes Q: can you see clouds? A: yes Q: can you see the sand on the beach? A: yes Q: was topole around?	A: yes Q: is the photo in color? A: yes C: is there any people? A: of C: is wrany birds? A: is in tight? A: is of its daytime Q: is is daytime Q: is is daytime Q: is is daytime Q: is is daytime? A: ope A: ope	A: yes Q: is the photo in color? A: yes C: is there any people? A: is C: is there any people? A: is C: is the photo in color? A: is in night? A: is A: is in high? A: is the sky visible? A: yes C: is the sky visible? A: yes C: are there many buildings? A: yes C: are there many buildings? A: yes C: are there treas! A: yes C: are there treas! C: is the photo in color? A: yes C: can you see clouds? A: yes C: can you see the sand on the beach? A: yes
Q: what cours is ue off(1) A: white Q: what cours is ue off(1) Q: what cours is ue off(1) a bird flying over the ocean beach at sunrise Q: do you see any boats? Q: any other birds? Q: any other birds? Q: any other birds? Q: any other birds? A: no Q: any jets or balloons? Q: do you see buildings? <		A: yes A: yes A: yes A: yes A: yes A: sys A: so A:	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is the bird brown? A: no Q: is there reses? A: yes Q: is the aky visible? Q: is the three namy buildings? A: yes Q: are the trees lush? A: yes Q: are the trees lush? A: yes Q: are the trees lush? A: yes Q: are not the cocolor? A: yes Q: are not the cocolor? A: yes Q: any hut color is the ocean? A: blue Q: any poople around? A: yes	A: yes () () () () () () () () () () () () ()	A: yes () () () () () () () () () () () () ()
a bird flying over the ocean beach at sunrise		A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: no Q: how many birds? A: 1 Q: is in the bird brown? A: no if is white Q: is in dury? A: yes Q: are there trees? A: yes Q: are there trees? A: yes Q: are there there show? A: point of them Q: are there then? A: yes Q: are there subs? A: yes Q: is the photo in color? A: yes Q: is the color is the ocean? A: yes Q: can you see clouds? A: yes Q: can you see the sand on the beach? A: yes Q: any people around? A: no	Q: is the photo in color? x; yes Q: is there any people? xi no Q: how many birds? xi no Q: how many birds? xi no Q: how many birds? xi no Q: is thre bird brown? xi no Q: is the bird brown? xi ros Q: are there remary buildings? xi yes Q: is the photo in color? xi yes Q: can you see clouds? Xi yes Q: what color is the ocean? Xi blue Q: can you see the sand on the beach? Xi yes Q: any people around? Xi no Q: what color is the bird?	A: yes Q: is the photo in color? A: yes Q: is there any people? A: is C: is there any people? A: is C: is any people Q: is the bird Q: are there trees? A: yes Q: are there anny buildings? A: yes Q: are the trees lush? A: yes Q: can you see clouds? A: yes Q: can you see the sand on the beach? A: yes Q: any people around? A: yes	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: no Q: how many birds? A: no Q: is the bird brown? A: no Q: is the bird brown? A: yes Q: are there trees? A: yes Q: are there shar? A: yes Q: early ou see clouds? A: yes Q: ou yes A: yes Q: yes Q: yes A: yes Q: yes
a bird flying over the ocean beach at sunrise beach at sunrise beach at sunrise beach at sunrise beach at sunrise beach at sunrise beach at sunrise control of the sunrise control of t		A: yes A: yes A: yes A: yes A: yes A: sys A: so A:	Q: is the photo in color? x; yes Q: is there any people? xi no Q: how many birds? xi no Q: how many birds? xi no Q: how many birds? xi no Q: is thre bird brown? xi no Q: is the bird brown? xi ros Q: are there remary buildings? xi yes Q: is the photo in color? xi yes Q: can you see clouds? Xi yes Q: what color is the ocean? Xi blue Q: can you see the sand on the beach? Xi yes Q: any people around? Xi no Q: what color is the bird?	A: yes A: yes A: yes A: the photo in color? A: no A:	A: yes Q: is the photo in color? Q: is the photo in color? Q: by yes Q: is there any people? A: no Q: how many birds? A: no Q: is the bird brown? A: no Q: is the bird brown? A: no Q: is the bird brown? A: yes Q: are there stres? A: yes Q: are there stres? A: yes Q: are there stres? A: yes Q: are there short? A: yes Q: are there short? A: yes Q: are the tres lush? A: yes Q: are the tres lush? A: yes Q: are the tres lush? A: yes Q: are the tres? A: yes Q: are the tres? A: yes Q: are the tres? A: yes Q: are the tres? A: yes Q: hat color is the ocean? A: yes Q: any people around? A: yes Q: any people around? A: no Q: what color is the bird?
beach at sunrise A: no A: no A: no Q: any other birds? Q: any other birds? Q: any other birds? Q: any other birds? A: 1 other A: no Q: any other birds? Q: any other birds? A: 1 other A: no A: no A: no Q: any other birds? Q: any other birds? Q: any other birds? Q: any other birds? Q: any jets or balloons? A: no A: no A: no A: no A: no Q: do you see buildings?	to some buildings	A: yes Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in inght? A: in Q: is in inght? A: in Q: is in inght? A: in Q: is in inght? A: yes Q: is in inght? A: yes Q: are there trees? A: yes Q: is the day visible? A: no Q: are there trees? A: yes Q: are there trees lush? A: yes Q: are the trees lush? A: yes Q: are the trees lush? A: yes Q: are holor is the ocean? Q: can you see clouds? A: yes Q: can you see clouds? A: yes Q: any people around? A: yes Q: any people around? A: white	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is the photo in color? A: yes Q: is the bird brown? A: no Q: is the bird brown? A: no Q: is the bird brown? A: yes Q: are there rese? A: yes Q: are there many buildings? Q: are there the rees lush? A: yes Q: ear the trees lush? A: yes Q: are the trees lush? A: yes Q: are not puous see clouds? A: yes Q: what color is the ocean? A: blue Q: any poople around? A: white	A: yes Q: is the photo in color? A: yes Q: is there any people? A: of the original of the original of the original A: no it is daytime Q: is the bird brown? A: no it is daytime Q: is the bird brown? A: no it is daytime Q: is the bird brown? A: original of the original Q: is the bird brown? A: yes A: ye	A: yes Q: is the photo in color? A: yes Q: is there any people? A: is pice of the photo in color? A: is pice of the photo in color? A: is in tight? A: is the bird brown? A: is the sky visible? A: is the sky visible? A: is the sky visible? A: is the there subs? A: is the the trees lush? A: is the the tool in color? A: is the the tool is the ocean? A: green by use che shand on the beach? A: is no Q: what color is the bird? A: white
Q: any outer onto:: A: no Q: any outer onto:: <	to some buildings	A: yes A: yes A: yes A: yes A: yes A: sys A: so A: so A: so A: no A:	Q: is the photo in color? x; yes Q: is there any people? xi no Q: is the bird brown? xi no Q: is it day? xi yes Q: are there trees? xi yes Q: are there many buildings? xi yes Q: is the photo in color? xi yes Q: can you see clouds? Xy yes Q: can you see the sand on the beach? Xy res Q: can you see the sand on the beach? Xy res Q: any people around? Xi no Q: what color is the bird? Xi no Q: why the color is the bird? Xi no	A: yes A: yes A: yes A: the photo in color? A: no A:	A: yes Q: is the photo in color? Q: is the photo in color? Q: is there any people? A: no Q: how many birds? A: no Q: is the bird brown? A: no Q: is the bird brown? A: no Q: is it day? A: yes Q: are there stres? A: yes Q: are there stres? A: yes Q: are there shub? A: yes Q: are there shub? A: yes Q: are the tres lush? A: yes Q: shub color is the ocean? A: yes Q: any people around? A: no Q: any people around? A: no Q: shub color is the bird? A: white Q: shy ta color is the bird? A: white A: whi
A: 1 other Q: any jets or balloons? A: no A: no A: no Q: any jets or balloons? A: no Q: any jets or balloons? Q: any jets or balloons? A: no Q: any jets or balloons? A: no A: no Q: do you see buildings? Q: do you see buildings? Q: do you see buildings?	to some buildings	A: yes A: yes A: yes A: yes A: yes A: no A: yes A:	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is the photo in color? A: yes Q: is the photo brown? A: no Q: is the bird brown? A: no Q: is the bird brown? A: yes Q: are there trees? A: yes Q: is the sky visible? A: o Q: are there many buildings? A: yes Q: are the trees lush? A: yes Q: is the photo in color? A: yes Q: is the photo in schor? A: yes Q: is the nour is the ocean? A: blue Q: any up see clouds? Y: yes A: what color is the ocean? A: blue Q: any up see the sand on the beach? A: yes Q: what color is the bird? A: white Q: do you see any boats?	A: yes A: no A:	A: yes () () () () () () () () () () () () ()
A: no A: no A: no A: no A: no A: no Q: do you see buildings?	to some buildings	A: yes A: yes A: yes A: yes A: yes A: sys A: so A: so A: so A: no A:	Q: is the photo in color? x; yes Q: is there any people? xi no Q: how many birds? xi no Q: how many birds? xi no Q: is the bird brown? xi no Q: at the bird brown? xi ros Q: are there reners? xi yes Q: are there many buildings? xi yes Q: are the trees lush? xi yes Q: can you see clouds? Xi yes Q: can you see clouds? Xi yes Q: can you see the sand on the beach? Xi yes Q: any people around? Xi no Q: what color is the bird? Xi no <t< td=""><td>A: yes A: yes A: yes A: the photo in color? A: to A: no A:</td><td>A: yes Q: is the photo in color? Q: is yes Q: is the photo in color? Q: is yes Q: is there any people? A: no Q: how many birds? A: no Q: is the bird brown? A: no Q: is the bird brown? A: no Q: is it day? A: yes Q: are there stres? A: yes Q: are there stres? A: yes Q: are there stres? A: yes Q: are there study? A: yes Q: are the tool in color? A: yes Q: are people around? A: no Q: any people around? A: no Q: what color is the bird? A: white Q: who color is the bird? A: no Q: any people around? A: no A: n</td></t<>	A: yes A: yes A: yes A: the photo in color? A: to A: no A:	A: yes Q: is the photo in color? Q: is yes Q: is the photo in color? Q: is yes Q: is there any people? A: no Q: how many birds? A: no Q: is the bird brown? A: no Q: is the bird brown? A: no Q: is it day? A: yes Q: are there stres? A: yes Q: are there stres? A: yes Q: are there stres? A: yes Q: are there study? A: yes Q: are the tool in color? A: yes Q: are people around? A: no Q: any people around? A: no Q: what color is the bird? A: white Q: who color is the bird? A: no Q: any people around? A: no A: n
A: no Q: do you see buildings? A: no	to some buildings	A: yes A: yes A: yes A: yes A: yes C: is there any people? A: no C: how many birds? A: no C: how how the experiment of t	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is the photo in color? A: yes Q: is the bird brown? A: no Q: is the bird brown? A: no Q: is the bird brown? A: yes Q: at draw there rese? A: yes Q: at here kay visible? A: no Q: are there many buildings? A: yes Q: are there many buildings? A: yes Q: are there many buildings? A: yes Q: are how use clouds? A: yes Q: are you see clouds? A: yes Q: any you see the sand on the beach? A: yes Q: white color is the bird? A: white Q: dy ous see any boats? A: no Q: any other birds? A: no	A: yes A: no A	A: yes () () () () () () () () () () () () ()
Q: do you see buildings?	to some buildings	A: yes A: yes A: yes A: yes A: yes A: yes A: sys A: so A:	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is the bird brown? A: no Q: is in tight? A: no Q: is the bird brown? Q: is the bird brown? Q: is the bird brown? Q: is the key visible? A: yes Q: are there trees? A: yes Q: are there many buildings? A: yes Q: are there subh? A: yes Q: are the trees lush? A: yes Q: any you see clouds? A: yes Q: whit color is the occan? A: blue Q: any you see the sand on the beach? A: yes Q: white white? A: white Q: why they be arrow boats? A: hue Q: any	A: yes A: no A	A: yes Q: is the photo in color? Q: is the search of the s
A: no A: no A: no	to some buildings	A: yes A: yes A: yes A: yes A: yes C: is there any people? A: no C: how many birds? A: no C: how how many birds? A: no C: how how how how many C: how how how how many C: how how how how many C: how how how how how many C: how	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is the photo in color? A: so Q: is the bird brown? A: no Q: at head brown? A: no Q: at the sky visible? A: no Q: are there many buildings? A: yes Q: are there many buildings? A: yes Q: are there many buildings? A: yes Q: are pose clouds? A: yes Q: expression was esclouds? A: yes Q: expression was use clouds? A: yes Q: expression see clouds? A: yes Q: what color is the occan? Q: what color is the bird? A: white Q: dy ous ace any boats? A: no	A: yes A: no A:	A: yes (); is the photo in color? A: yes (); is the photo in color? A: yes (); is there any people? A: no (); how many birds? A: no (); is the bird brown? A: no (); is the bird brown? A: no (); is the bird brown? A: yes (); is the bird brown? A: yes (); is the sky visible? A: yes (); is the sky visible? A: yes (); is the sky visible? A: yes (); is the photo in color? A: yes (); is the photo in the bird? A: yes (); is the color is the ocean? A: yes (); is the color is the bird? A: yes (); any pople around? A: no (); any other birds? A: no (); any is or balloons? A: no (); any is or balloons? A: no (); any is or balloons? (); any color balloons? (); any color balloons? (); any color balloons? (); any color
	to some buildings	A: yes A: yes A: yes A: yes A: yes C: is there any people? A: no C: how many birds? A: no C: how how many birds? A: no C: how how how how? A: no C: how how how how? A: no C: how how how how? A: no C: how	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in tinght? A: no Q: is in tinght? A: no Q: is in tinght? A: no Q: is the bird brown? A: no Q: is the bird brown? A: no Q: is the key visible? A: ros Q: is the key visible? A: ros Q: are there many buildings? A: yes Q: are there many buildings? A: yes Q: are the trees lush? A: yes Q: are the trees lush? A: yes Q: are not on you see clouds? A: yes Q: any up sepile around? A: no Q: any up set he around? A: no Q: any up there? A: no Q: any up ther bird?? A: no Q: any up ther bird?? A: no Q: any up ther bird?? A: no </td <td>A: yes A: no A:</td> <td>A: yes (); is the photo in color? A: yes (); is the photo in color? A: yes (); is there any people? A: no (); how many birds? A: no (); is the bird brown? A: no (); is the bird brown? A: no (); is the bird brown? A: yes (); is the bird brown? A: yes (); is the sky visible? A: yes (); is the sky visible? A: yes (); is the sky visible? A: yes (); is the photo in color? A: yes (); is the photo in the bird? A: yes (); is the color is the ocean? A: yes (); is the color is the bird? A: yes (); is the color is the bird? (); white: (); dy us see any boats? A: no (); my people around? A: no (); my istor balloons? A: no (); my istor balloons? (); is no (); my people around? (); no (); my istor balloons? (); no (); my people around? (); hit is no ()</td>	A: yes A: no A:	A: yes (); is the photo in color? A: yes (); is the photo in color? A: yes (); is there any people? A: no (); how many birds? A: no (); is the bird brown? A: no (); is the bird brown? A: no (); is the bird brown? A: yes (); is the bird brown? A: yes (); is the sky visible? A: yes (); is the sky visible? A: yes (); is the sky visible? A: yes (); is the photo in color? A: yes (); is the photo in the bird? A: yes (); is the color is the ocean? A: yes (); is the color is the bird? A: yes (); is the color is the bird? (); white: (); dy us see any boats? A: no (); my people around? A: no (); my istor balloons? A: no (); my istor balloons? (); is no (); my people around? (); no (); my istor balloons? (); no (); my people around? (); hit is no ()
	to some buildings	A: yes A: yes A: yes A: yes A: yes A: sys A: so A: so A: so A: no A:	Q: is the photo in color? A: yes Q: is there any people? A: no Q: how many birds? A: 1 Q: is in tinght? A: no Q: is in tinght? A: no Q: is in tinght? A: no Q: is the bird brown? A: no Q: is the bird brown? A: no Q: is the key visible? A: ros Q: is the key visible? A: ros Q: are there many buildings? A: yes Q: are there many buildings? A: yes Q: are the trees lush? A: yes Q: are the trees lush? A: yes Q: are not on you see clouds? A: yes Q: any up sepile around? A: no Q: any up set he around? A: no Q: any up there? A: no Q: any up ther bird?? A: no Q: any up ther bird?? A: no Q: any up ther bird?? A: no </td <td>A: yes A: yes A: yes A: the photo in color? A: so A: no A:</td> <td>A: yes Q: is the photo in color? A: yes A: no Q: is the carry people? A: i Point of the photo in color? A: no Q: is the bird brown? A: no Q: is the bird brown? A: no Q: is the bird brown? A: yes Q: are there stres? A: yes Q: are the tres lush? A: yes Q: are the tres? A: no Q: start the tres? A: no Q: start the tres? A: no Q: are yes the stres? A: no A:</td>	A: yes A: yes A: yes A: the photo in color? A: so A: no A:	A: yes Q: is the photo in color? A: yes A: no Q: is the carry people? A: i Point of the photo in color? A: no Q: is the bird brown? A: no Q: is the bird brown? A: no Q: is the bird brown? A: yes Q: are there stres? A: yes Q: are the tres lush? A: yes Q: are the tres? A: no Q: start the tres? A: no Q: start the tres? A: no Q: are yes the stres? A: no A:

Fig. 7: Qualitative samples for three model variants – ViLBERT w/ CC + VQA (called 'Base'), Base + CE, and Base + CE + NSP.