Weakly-Supervised Generation and Grounding of Visual Descriptions with Conditional Generative Models

Effrosyni Mavroudi and René Vidal
Mathematical Institute for Data Science, Dept. of Biomedical Engineering, Johns Hopkins University
{emavrou1, rvidal}@jhu.edu

Abstract

Given weak supervision from image- or video-caption pairs, we address the problem of grounding (localizing) each object word of a ground-truth or generated sentence describing a visual input. Recent weakly-supervised approaches leverage region proposals and ground words based on the region attention coefficients of captioning models. To predict each next word in the sentence they attend over regions using a summary of the previous words as a query, and then ground the word by selecting the most attended regions. However, this leads to sub-optimal grounding, since attention coefficients are computed without taking into account the word that needs to be localized. To address this shortcoming, we propose a novel Grounded Visual Description Conditional Variational Autoencoder (GVD-CVAE) and leverage its latent variables for grounding. In particular, we introduce a discrete random variable that models each word-to-region alignment, and learn its approximate posterior distribution given the full sentence. Experiments on challenging image and video datasets (Flickr30k Entities, YouCook2, ActivityNet Entities) validate the effectiveness of our conditional generative model, showing that it can substantially outperform soft-attention-based baselines in grounding.

1. Introduction

Linking words to visual regions provides a fine-grained bridge between vision and language modalities and is a fundamental block of many applications, such as human-robot interaction [57, 60], visual question answering [27, 61], and even unsupervised neural machine translation [58]. Thus, visual grounding has become a prominent research area at the intersection of vision and language [12, 16, 29, 51]. Training visual grounding systems typically requires annotations of textual descriptions combined with bounding boxes for each groundable word (e.g., object nouns). Since constructing datasets with such fine-grained bounding box annotations is rather time-consuming and costly, we focus

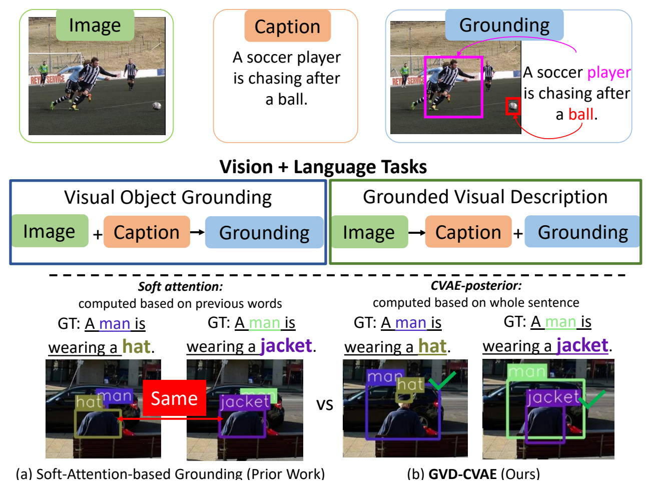

Figure 1. **Our proposed framework jointly models visual descriptions and word-to-region alignments** conditioned on an input image (or video) and region proposals. Without using any bounding box annotations during training, it can tackle two tasks: Visual Object Grounding and Grounded Visual Description. Unlike prior work [74] that leverages soft attention for grounding and always predicts the same region for two words given the same visual input and partial caption context, our model can ground words by taking into account the full ground-truth or generated sentence.

on weakly-supervised training of visual grounding systems, which require only image-caption pairs for training. In particular, we consider two tasks, as illustrated in Fig. 1: (1) *Weakly-Supervised Visual Object Grounding (WS-VOG)*, where given an input image (or video) and its visual description, the goal is to *localize* the referred semantic entities in the visual input, and (2) *Weakly-Supervised Grounded Visual Description (WS-GVD)*, where given an input image (or video), we must jointly *generate* a natural language description and *localize* the generated words.

Most prior work has focused on learning how to align words with regions by learning how to correctly match images and videos to sentences [8, 26, 56, 65]. However, these matching-based approaches can only tackle the first task (*WS-VOG*), and cannot generate grounded visual descriptions. On the other hand, captioning-based approaches [40, 74] aim to learn how to ground words by learning how to generate captions based on region proposals, thus they can tackle both tasks. For example, the GVD captioning-based model [74] grounds words by using the region attention mechanism of a discriminative, encoder-decoder captioning model to select regions with maximum attention coefficients. Nonetheless, exploiting soft attention as a grounding mechanism suffers from two major limitations. First, despite being an effective, end-to-end learnable mechanism for summarizing relevant context, attention is not explicitly encouraged to capture meaningful alignments and can result in poor grounding [36], unless it is supervised. Second, each word is generated using attention coefficients computed from a query that summarizes the previously generated words. Hence, the coefficients do not take into account the word to be grounded. For example, consider grounding the words *'hat'* and *'jacket'* given the sentences *"A man is wearing a hat"* and *"A man is wearing a jacket"*, respectively. As shown in Fig. 1, existing attention-based grounding approaches wrongly predict the same box for 'hat' and 'jacket', since the partial caption is the same.

To overcome these limitations, we propose a conditional generative model for the joint probability distribution of sentences and latent word-to-region alignments given an input image (or video) and a set of region proposals. That is, we account for the lack of grounding annotations by introducing discrete latent variables that model word-to-region alignments. We parameterize our model with state-of-the-art visual encoders, language decoders and attention modules, and leverage Amortized Variational Inference [30, 59] to learn its parameters. The resulting Grounded Visual Description Conditional Variational Autoencoder (GVD-CVAE) allows us to both generate sentences and also infer the latent word-to-region alignments based on the *whole sentence, including the word to be grounded*. Hence, it can correctly ground the *hat* in the motivating example.

In summary, this work makes three key contributions. First, we introduce the GVD-CVAE, a novel conditional generative model of visual descriptions with a sequential discrete latent space and attention-based parameterization of the prior and approximate posterior alignment distributions. Second, we propose a training objective that encourages our model to learn latent variables that capture meaningful word-to-region alignments. Third, we evaluate our method on three challenging image and video datasets and demonstrate that both our "prior" and "approximate posterior" alignment distributions improve upon soft attention. This leads to a 12% absolute improvement in *WS-VOG* on Flickr30k Entities. Our model also achieves state-of-the-art or competitive grounding and captioning performance compared with a diverse family of state-of-the-art methods that are tailored to *WS-VOG* or *WS-GVD*.

2. Related Work

Grounded Visual Description. Developing models that can both generate a sentence and link the generated words to visual regions is a nascent research area, motivated by a need for more trustworthy and interpretable captioning models [24,36,50]. Such models can be seen as an evolution of early image auto-annotation methods [7], methods for generating visually grounded storylines [20], or methods for generating descriptions with grounded and co-referenced people [52]. Zhou et al. [74] ground words by leveraging the region attention coefficients of an attention-based captioning models. However, in contrast to prior work on phrase grounding that computes attention using the whole phrase as query [51], the region attention in [74] is computed based on previous words (partially generated sentence), and it is thus agnostic to the word being grounded.

A recent line of work has attempted to mitigate this issue. Ma et al. [40] propose a cyclical training regime for *WS-GVD* of images and videos that involves two attention mechanisms: one based on the partial caption and another based on the groundable word. By forcing the words generated using these two attention mechanisms to match the ground-truth words, the mechanisms are implicitly regularized to produce similar attention weights during training. Other approaches explicitly supervise the region attention during training on image-caption pairs, either by using attention coefficients based on future relevant words [37], or by leveraging the word-to-region alignments of a separately trained image-to-text matching model [77]. In summary, a common thread in prior work is the usage of a regular region attention module of an UpDown [2] captioning model for grounding, which is regularized *only during training* based on auxilliary models or attention mechanisms. In contrast, inspired by discrete latent-variable models for image captioning/neural machine translation [13, 45, 54, 66], our key innovation is to treat word-to-region alignments as discrete latent variables in a grounded visual description CVAE model and exploit the prior or approximate posterior alignment distributions to infer the latent word-to-region alignments. This enables us to consider the past, future and current words for localizing each object word in the input image or video *during testing*.

Visual Object Grounding. Grounding words (rather than whole sentences [71] or phrases [21, 65]) in images and videos is an active research field in the intersection of vision and language. Early attempts for weakly-supervised visual grounding given textual descriptions of images and videos relied on graphical models [47, 68]. Powered by advances in region proposal generation, a large group of recent methods [11, 29] cast the task as a *Multiple Instance Learning* (MIL) problem. These methods define an image-sentence matching score determined by word-to-region alignments and learn how to correctly match images to sentences using ranking losses. Such methods have also been extended to videos [26, 56, 75] with frame-sentence matching scores and mechanisms to account for missing objects. However, these MIL-based methods cannot both generate sentences and ground objects. This limitation is lifted by the *captioning-based* GVD-Grd method [74], which grounds each word based on region attention coefficients, computed with the previous words as query, combined with region-to-class similarity coefficients. These are obtained by transferring object class knowledge from external datasets. In this work, we also use captioning as a downstream task, but we localize words with the distributions of a conditional generative model, leveraging the full sentence context.

Joint Vision-Language Representation Learning. Inspired by advances in pretrained NLP models [14], researchers have also started to use large-scale vision-text corpora to learn cross-modal vision-language representations. There exist Transformer-based models [35, 38] that are also trained using only pairs of images with object proposals and associated textual descriptions. However, instead of focusing on learning task-agnostic, visiolinguistic representations using large-scale corpora to facilitate downstream tasks, we are interested in training visual grounding systems on small-scale datasets. Importantly, we rely on text as weak supervision for learning how to ground without bounding box annotations directly on the target dataset. Instead, these pretrained models require finetuning on a smaller, fully-annotated dataset to tackle downstream tasks such as referring expression grounding [38].

Modeling Sequential Data with Variational Autoencoders. Our proposed CVAE-based captioning model is also related to regular or Conditional VAEs that are developed for modeling sequential data in NLP applications. In particular, VAEs with *sequences of latent variables* [3, 9, 10, 18, 53, 69] instead of a single latent variable driving the whole sequential generation process [5, 43, 63, 72] are more closely related to our work. However, the majority of those have non-interpretable, continuous latent variables, unlike our discrete latent word-to-region alignments. A notable exception is the approach of Graber et al. [19] that uses sequential discrete variables to model interactions between entities in interacting systems. Still, all these works share the same goal of modeling the likelihood of sequential data, while we propose exploiting the latent variables for grounding. To this end, we need to avoid training an inference model that produces posteriors almost identical to the prior, thus ignoring the word to be grounded. Researchers are actively exploring various techniques to mitigate this *posterior collapse* issue by modifying: the training objective [1, 17, 34, 42, 48, 55], the training procedure [22] or the decoder architecture [15]. Similarly, we propose controlling the relative factor between sentence reconstruction term and the prior regularization term [1, 6, 55].

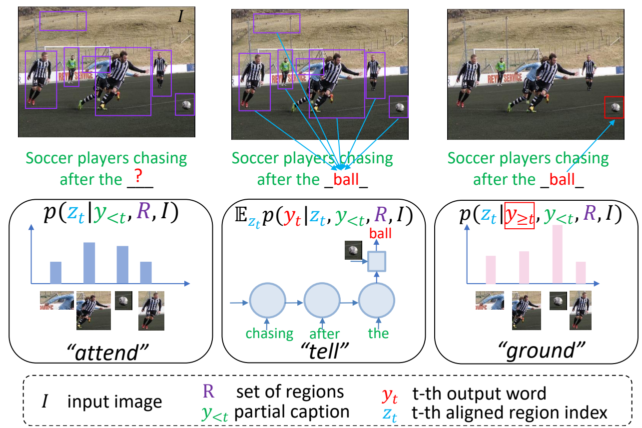

Figure 2. **We propose a deep conditional generative model of visual descriptions** that models each word-to-region alignment with a discrete latent variable $\mathbf{z}_t$. It is able to *attend* over the region proposals in an input image (or video), *tell* what it shows by marginalizing out the latent word-to-region alignments from the joint distribution and *ground* each word by leveraging the learned approximate posterior word-to-region alignment distribution.

3. Method

3.1. Problem Formulation

Let Y denote a visual description of a given visual input I (*i.e.*, an image or video). We represent $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$ as a sequence of T words from a vocabulary $\mathcal{V}$, where $\mathbf{y}_t$ is the one-hot encoding of the t-th word, *i.e.*, $\mathbf{y}_t \in \{0,1\}^{|\mathcal{V}|}$ and $\|\mathbf{y}_t\| = 1$. In the *VOG* task, the goal is to ground words in ground-truth descriptions of a visual input, *i.e.*, we are interested in localizing each mentioned groundable word with a bounding box $\widehat{\mathbf{b}}_t$. In the *GVD* task, the goal is to both generate a visual description $\widehat{Y}$ and localize each generated groundable word $\widehat{\mathbf{y}}_t$ with a bounding box $\widehat{\mathbf{b}}_t$.

In this work, we propose to design a model that can tackle both tasks in both the image and video domains, and can be trained with weak supervision in the form of aligned visual input and and visual description pairs $\{(I^{(n)}, Y^{(n)})\}_{n=1}^N$. To achieve this, we treat the problem of grounding as a problem of word-to-region alignment by leveraging M candidate region proposals $R = \{\mathbf{r}_m\}_{m=1}^M$ obtained by an off-the-shelf object detector [23]. Then, the localization problem is reduced to identifying the variable $\mathbf{z}_t \in \{0,1\}^M$ with $\|\mathbf{z}_t\| = 1$, which denotes which region corresponds to the t-th word. Our key idea is to model word-to-region alignments as latent variables in a deep conditional generative model. To this end, we propose a novel Grounded Visual Description Conditional Variational Autoencoder (GVD-CVAE). As illustrated in Fig. 2, learning such a model allows us to leverage the posterior distribution of word-to-region alignments for grounding words *based on the entire sentence*, unlike attention-based grounding.

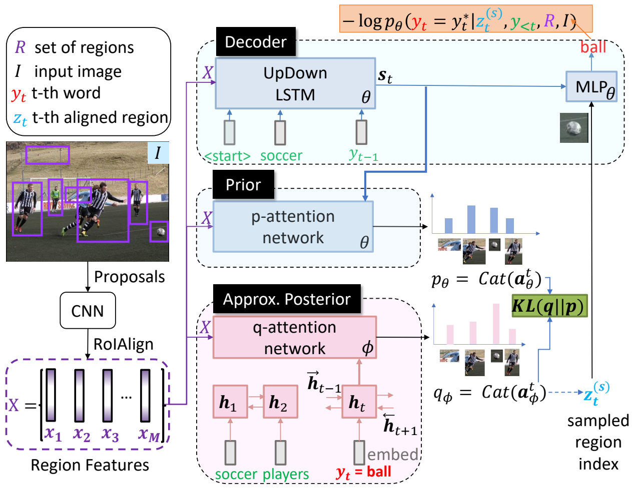

Figure 3. **Our proposed GVD-CVAE architecture**. The input image and proposals are fed through a *visual encoder* to produce region embeddings. The *prior word-to-region alignment* is computed as a function of only the previous words, while the *approximate posterior* is computed as a function of the full sentence. During training, a region is sampled from the approximate posterior and is fed to the *language decoder* that predicts the next word.

3.2. Attention-based Conditional Variational Autoencoder for Grounded Visual Description

Let $Z = \{\mathbf{z}_1, \dots, \mathbf{z}_T\}$ be the sequence of latent variables corresponding to alignments between words and regions, where $\mathbf{z}_t \in \{0,1\}^M$ is a binary discrete random variable with $z_{t,i} = 1$ when the i-th region proposal corresponds to the t-th word $\mathbf{y}_t$. The joint conditional distribution $p_\theta(Y, Z \mid R, I)$ of a caption Y and sequence of alignments Z, given the input video (or image) I and candidate regions R can be factorized in an autoregressive manner:

$$\prod_{t=1}^{T} p_\theta(\mathbf{y}_t \mid \mathbf{y}_{<t}, \mathbf{z}_{\leq t}, R, I) p_\theta(\mathbf{z}_t \mid \mathbf{y}_{<t}, \mathbf{z}_{<t}, R, I), \quad (1)$$

where $\mathbf{y}_{<t} = \mathbf{y}_{1:t-1}$ is the partial caption up to word $t-1$, and similarly $\mathbf{z}_{<t}$ denotes the sequence of word-to-region alignments up until word $t-1$. We can simplify this joint distribution by making two assumptions: (a) the t-th word depends only on the region $\mathbf{z}_t$ given the partial caption $\mathbf{y}_{<t}$, and (b) the region-to-word alignments $\mathbf{z}_t$ for each word are conditionally independent of each other given the partial caption. Hence, our joint probability distribution becomes:

$$p_\theta(Y, Z|R, I) = \prod_{t=1}^{T} \overbrace{p_\theta(\mathbf{y}_t|\mathbf{y}_{<t}, \mathbf{z}_t, R, I)}^{\text{language decoder}} \overbrace{p_\theta(\mathbf{z}_t|\mathbf{y}_{<t}, R, I)}^{\text{region prior}}. \quad (2)$$

Next, we describe how we parameterize our conditional generative model with deep networks whose trainable weights are denoted by θ, as illustrated in Fig. 3.

Visual Encoder. Images are encoded using a pretrained CNN model with RoI-pooling operations and trainable linear projections [74]. The encoder captures global visual context in the form of a coarse image-level feature vector, $\mathbf{v}$, as well as fine-grained grid features $F = \{\mathbf{f}_l\}_{l=1}^{L}$, where l indexes the feature map spatial grid. It also generates grounding-aware region representations $X = \{\mathbf{x}_i\}_{i=1}^{M}$, where the representation $\mathbf{x}_i$ of each region encodes information about appearance, position and object class knowledge [74] transferred from an object detector [23] trained on an external dataset [31]. Videos are also encoded to a global video feature $\mathbf{v}$, a sequence of frame-level features $F = \{\mathbf{f}_l\}_{l=1}^{L}$, where l indexes the frames, and grounding-aware region representations $X = \{\mathbf{x}_i\}_{i=1}^{M}$, but using different network architectures, as detailed in the appendix.

Language Decoder. The decoder $p_\theta(\mathbf{y}_t|\mathbf{y}_{<t}, \mathbf{z}_t, R, I) = \mathrm{Cat}(g_\theta(\mathbf{s}_t, \mathbf{z}_t, X))$ is a categorical distribution over words in the vocabulary given the partial caption $\mathbf{y}_{<t}$, the word-to-region alignments $\mathbf{z}_t$, the regions R, and the visual input I. We parameterize this distribution with a shallow network

$$g_\theta(\mathbf{s}_t, \mathbf{z}_t, X) = \mathrm{softmax}(W_c \tanh(W_p \left[\mathbf{s}_t; \sum_{i=1}^{M} z_{t,i}\mathbf{x}_i\right])), \quad (3)$$

whose inputs are: (a) the state $\mathbf{s}_t \in \mathbb{R}^d$ of a language model that summarizes $\mathbf{y}_{<t}$, R and I, and (b) the aligned region feature $\sum_{i=1}^{M} z_{t,i}\mathbf{x}_i \in \mathbb{R}^d$, where $[\cdot;\cdot]$ denotes concatenation, and $W_c \in \mathbb{R}^{d \times d}, W_p \in \mathbb{R}^{d \times 2d}$ are learnable weights.

Although $\mathbf{s}_t$ can be chosen as the state of any standard language model [2, 76], we follow prior work on grounded visual description [37, 40, 74, 77] and adopt a variant of the UpDown [2] LSTM model. This language model is composed of a word embedding layer (emb) and two LSTM [25] layers with hidden states $\mathbf{u}_t$ and $\mathbf{s}_t$, respectively. It also uses an additive attention mechanism [4] $f_\theta(\mathbf{u}_t, F) = \mathrm{softmax}(\mathbf{w}_f^T \tanh(W_f[\mathbf{u}_t; \mathbf{f}_l]))$ over holistic visual features F, where $\mathbf{w}_f, W_f$ are learnable attention weights. Region features X are also summarized with another additive attention mechanism $k_\theta(\mathbf{u}_t, X)$.

$$\mathbf{u}_t = \mathrm{RNN}_\theta^1(\mathbf{u}_{t-1}, [\mathbf{v}; \mathrm{emb}(\mathbf{y}_{t-1})]) \quad (4)$$

$$\mathbf{s}_t = \mathrm{RNN}_\theta^2\left(\mathbf{s}_{t-1}, \left[\sum_{l=1}^{L} f_\theta^{(l)}(\mathbf{u}_t, F)\mathbf{f}_l; \sum_{i=1}^{M} k_\theta^{(i)}(\mathbf{u}_t, X)\mathbf{x}_i; \mathbf{u}_t\right]\right). \quad (5)$$

Prior Model. The prior distribution, $p_\theta(\mathbf{z}_t|\mathbf{y}_{<t}, R, I)$, is a categorical distribution over possible word-to-region alignments. We choose to parameterize it with an additive attention mechanism [4] that computes region attention coefficients $\alpha_\theta(\mathbf{s}_t, X) \in \mathbb{R}^M$ using as a query the top LSTM

state $\mathbf{s}_t$ that summarizes the partial caption and visual input:

$$\mathbf{z}_t \mid \mathbf{y}_{<t}, R, I \sim \mathrm{Cat}(\alpha_\theta(\mathbf{s}_t, X)). \tag{6}$$

Variational Posterior. To learn the parameters of our conditional generative model we leverage Amortized Variational Inference (AVI). Therefore, our model becomes a CVAE [59] with sequential discrete latent space and sentences as observations. In the CVAE framework, a variational distribution $q_\phi(Z|Y, R, I)$ is introduced to approximate the true posterior and is parameterized via a neural network with weights ϕ, also known as the "inference network". We choose to approximate the true posterior with the following approximate posterior:

$$q_\phi(Z \mid Y, R, I) = \prod_{t=1}^{T} q_\phi(\mathbf{z}_t \mid Y, R, I). \tag{7}$$

Then, we model the approximate posterior distribution of each word-to-region alignment as a categorical distribution that is parameterized by the attention coefficients $\alpha_\phi(\mathbf{h}_t, X) \in \mathbb{R}^M$ obtained via another attention network, implemented as additive attention [4] or general dot-product attention [39]:

$$\mathbf{z}_t \mid Y, R, I \sim \mathrm{Cat}(\alpha_\phi(\mathbf{h}_t, X)). \tag{8}$$

In this case, the attention query $\mathbf{h}_t \in \mathbb{R}^d$ summarizes the whole sentence. It is obtained by summing the forward and backward states of a BiLSTM network, whose inputs consist of the global feature $\mathbf{v}$ and ground-truth word $\mathbf{y}_t$ at each timestep. Optionally, we can augment the unnormalized attention coefficients $\tilde{\alpha}_\phi(\mathbf{h}_t, X)$ for the object words in the input sentence with transferred object class knowledge:

$$\begin{aligned}\mathbf{z}_t \mid Y, R, I \sim \mathrm{Cat}(\mathrm{softmax}(&\tilde{\alpha}_\phi(\mathbf{h}_t, X)+ \\ &+ \gamma\omega_t(\mathbf{w}_{c_t}^T\mathbf{o} + \mathbf{1}b_{c_t}))),\end{aligned} \tag{9}$$

where $\mathbf{w}_{c_t} \in \mathbb{R}^{d_0}, b_{c_t} \in \mathbb{R}$ are trainable weights, initialized with the pretrained object classifier for the external dataset's object class c_t that is closest to the object word $\mathbf{y}_t$, $\mathbf{o} \in \mathbb{R}^{d_o \times M}$ are region object features, $\mathbf{1} \in \mathbb{R}^M$ is a vector of all ones, and ω_t is a binary word mask with $\omega_t = 1$ denoting a groundable word. The hyperparameter $\gamma \in \{0, 1\}$ controls whether this transferred knowledge will be used or not.

Training. During training, we assume we are given N i.i.d. pairs of visual inputs and their visual descriptions, without grounding supervision. To train our Grounded Visual Description CVAE (GVD-CVAE), we minimize the following loss over the parameters θ and ϕ (omitting the conditioning of all distributions on $I^{(n)}$ for readability) :

$$\mathcal{L} = \frac{1}{N}\sum_{n,t} \lambda\mathcal{L}_{CVAE}(n,t) + (1-\lambda)\mathcal{L}_{CE}(n,t) \tag{10}$$

where $\mathcal{L}_{CE} = -\log p_\theta(\mathbf{y}_t^{(n)} \mid \mathop{\mathbb{E}}_{\mathbf{z}_t \sim p_\theta}[\mathbf{z}_t], \mathbf{y}_{<t}^{(n)}, R^{(n)})$ and

$$\begin{aligned}\mathcal{L}_{CVAE}(n,t) = &\mathop{\mathbb{E}}_{\mathbf{z}_t \sim q_\phi}\left[-\log p_\theta(\mathbf{y}_t^{(n)} \mid \mathbf{y}_{<t}^{(n)}, \mathbf{z}_t, R^{(n)})\right] \\ &+ \beta\mathrm{KL}\left(q_\phi(\mathbf{z}_t \mid Y^{(n)}, R^{(n)}) \mid\mid p_\theta(\mathbf{z}_t \mid \mathbf{y}_{<t}^{(n)}, R^{(n)})\right).\end{aligned} \tag{11}$$

For $\lambda = \beta = 1$, we recover the negative of the Evidence Lower Bound Objective (ELBO) for our factorization of the joint probability distribution and our choice of the approximate posterior. Similar to prior work in generative modeling [6], we observe that optimizing the ELBO often results in an inference model that produces approximate posteriors almost identical to the prior. To mitigate this issue, we reweight the KL loss term with a scalar factor β. We found that gradually increasing β up to a value $\beta_{clip} < 1$ during training or using a PI-controller [55] to reach a desired KL divergence value are effective for training our GVD-CVAE. Moreover, we experimentally observed that optimizing the CVAE loss jointly with a cross-entropy word prediction loss ($\lambda = 0.5$), that is applied on word predictions obtained based on the p-attention-based weighted sum of region features, further facilitates training. More details about training with Gumbel-Softmax [28, 41] samples and about approximate inference with our model are included in the appendix.

4. Experiments

4.1. Datasets, Metrics and Implementation Details

Flickr30k Entities (F30k) is a large-scale image dataset, originally annotated with phrase-to-region alignments [46]. To evaluate our results on object grounding (rather than phrase grounding), we follow the setup from Zhou et al. [74] to convert each noun phrase (e.g. *her brown hat*) associated with each bounding box to a single groundable object, such as *hat*. This results in $|\mathcal{V}_o| = 480$ groundable words out of the $|\mathcal{V}| = 8639$ words comprising the vocabulary. We use the standard dataset split with 29k/1k/1k images in the training, validation and testing sets, respectively.

ActivityNet Entities (ANet) is a large-scale video dataset, containing $52k$ video segments annotated with a caption each. Following the original setup [74], we use a vocabulary of 4905 words, 431 of which are groundable. Each groundable word in a sentence is associated with a bounding box in a frame of the video where it can be clearly observed. Since annotations for the testing set are not public and the evaluation server is closed at the time of submission, we follow [64] and report results on the validation set.

YouCook2-BB is a video dataset containing YouTube cooking videos with video segments paired with captions and bounding box annotations [75] at 1 fps for 67 object classes. We use the same training/validation/test split as in [56].

Table 1. **Comparison of grounding performance between the GVD and GVD-Grd models (baselines) and the GVD-CVAE on the validation sets of F30k and ANet**. We report the box accuracy metric for evaluating grounding given ground-truth sentences and the $F1_{all}$ metric for evaluating grounding of object words in generated sentences. GVD-CVAE-p (GVD-CVAE-q) denotes using our learned prior (approximate posterior) alignment distribution for grounding.

Dataset	Method	Box Acc.	$F1_{all}$
F30k (Image)	GVD [74]	22.0	4.4
	GVD-Grd [74]	25.9	4.4
	GVD-CVAE-p (Ours)	29.6	6.2
	GVD-CVAE-q (Ours)	**33.4**	**7.3**
ANet (Video)	GVD [74]	14.9	3.7
	GVD-Grd [74]	21.3	3.7
	GVD-CVAE-p (Ours)	19.4	4.8
	GVD-CVAE-q (Ours)	**24.2**	**6.1**

Metrics. Performance for *WS-VOG* is measured with Box Accuracy [56, 74, 75], which computes the percentage of correctly localized words of an object class. A word is considered to be correctly localized when its predicted box has more than 0.5 Intersection-over-Union (IoU) with ground-truth boxes. Metrics for *WS-GVD* evaluate both grounding and captioning capabilities. We adopt the $F1_{all}$ and $F1_{loc}$ grounding metrics [74] for evaluating grounding on generated sentences, and standard language evaluation metrics, such as Bleu [44], METEOR [32], CIDEr [62], and SPICE [2], for evaluating generated sentences. In $F1_{all}$, a region prediction is considered correct if the object word is both correctly predicted and localized, while $F1_{loc}$ only considers correctly predicted object words.

Implementation details. For the F30k and Anet datasets, our GVD-CVAE receives as inputs the region proposals, region features and image/video global features from Zhou et al. [74], with 100 region proposals per frame/image. For YouCook2, we use 20 region proposals and the features extracted by Shi et al. [56]. Hyperparameters such as learning rate, β_{clip}, attention mechanisms, number of samples, are chosen based on the validation sets of F30k and YouCook2. For evaluating on the ANet validation set, we train a model with hyperparameters selected based on the F30k validation set. All other hyperparameters, such as layer sizes, are in general adopted from prior work [56, 74]. Additional training and implementation details are included in the appendix.

4.2. Baselines and Ablation Studies

(1) Are the regions localized via our learned word-to-region alignment distributions better than those localized via soft-attention-based baselines? Our baseline is the attention-based encoder-decoder GVD captioning model, trained with teacher-forcing language generation cross-entropy loss. We ensure that our GVD-CVAE exactly mirrors the inputs and the visual encoder/language decoder modules of this baseline model. Baseline object grounding is performed either by (a) selecting the region with maximum region attention coefficient $k_\theta^{(i)}(\mathbf{u}_t, X)$ (GVD [74]) given the partial caption $\mathbf{y}_{<t}$, or (b) by combining the attention coefficients with region-to-class similarity scores based on the word $\mathbf{y}_t$ to be grounded for the *VOG* task (GVD-Grd [74]). In Table 1, we compare our GVD-CVAE's ability to ground objects in ground-truth or generated sentences with these two powerful, discriminative baselines. We observe that even grounding based on our learned *prior* word-to-region alignment distribution (GVD-CVAE-p) improves upon the soft-attention baseline by a significant margin in both benchmarks and tasks (*e.g.*, it improves Box Accuracy from 22% to 29% on F30k), despite similarly capturing only the history of previous words. The reason for this improvement is that our prior distribution is encouraged during training to "look ahead" when sampling a region to generate a word, by mimicking the approximate posterior alignment distribution which has access to future words. Using the latter for grounding conditioned on the full sentence further improves results (from 29.6% to 33.4%), verifying our intuition that leveraging the word to be grounded in its language context can help us better localize the word. Additionally, it outperforms the GVD-Grd discriminative baseline which also takes into account the word to be grounded, demonstrating the benefits of our conditional generative modeling.

Table 2. **Ablation analysis of the decoder and inference model design** on the F30k validation set. Types of UpDown [2] model attention: *Grid*: over grid features, *Reg.*: over region features, *Both*: both attention mechanisms. *Obj. Cls.* denotes inference model with transferred object class knowledge ($\gamma = 1$).

Decoder	Approximate Posterior		Box Acc.		$F1_{all}$	
	Cond.	Obj. Cls.	p	q	p	q
UpDown (Both)	$\mathbf{z}_t \vert \mathbf{y}_{\leq T}$	✓	29.6	33.4	6.2	7.3
UpDown (Both)	$\mathbf{z}_t \vert \mathbf{y}_{\leq T}$	✗	25.1	32.3	5.4	7.0
UpDown (Both)	$\mathbf{z}_t \vert \mathbf{y}_{\leq t}$	✗	26.3	31.4	6.0	7.4
UpDown (Grid)	$\mathbf{z}_t \vert \mathbf{y}_{\leq T}$	✓	**30.7**	34.4	**7.3**	7.2
UpDown (Reg.)	$\mathbf{z}_t \vert \mathbf{y}_{\leq T}$	✓	26.6	33.0	5.8	6.3
LSTM	$\mathbf{z}_t \vert \mathbf{y}_{\leq T}$	✓	30.2	**34.8**	6.9	**7.5**

(2) How does the choice of the language model and approximate posterior affect grounding performance? Table 2 demonstrates the grounding performance obtained with different design choices. First, results suggest that taking the full sentence into account via a BiLSTM ($q(z_t|y_{\leq T})$) leads to better *VOG* grounding compared to only seeing the sentence up to the current word $y_{\leq t}$ with an LSTM (e.g., improving Box Acc. from 31.4% to 32.3%)

Table 3. **Impact of various training objectives on weakly-supervised object grounding**. Performance measured via Box accuracy (%) on the F30k validation set.

Training objective	CVAE-p	CVAE-q
ELBO	3.29	3.16
CE + ELBO	25.22	23.99
CE + ELBO + β anneal	26.07	25.61
CE + ELBO + β anneal + clip	26.31	28.88
CE + ELBO + PI Controller	**29.27**	**31.71**

(rows 2-3). Another observation is that explicitly adding transferred information about object class distributions in the inference model ($\gamma = 1$) improves grounding given ground-truth sentences (Box Accuracy) with both the prior and approximate posterior distributions (rows 1-2). This further demonstrates that knowledge from the inference model is distilled to the prior during training via the KL loss, resulting in a model that is looking at better localized regions *while generating* descriptions based on the prior and decoder modules. Interestingly, results suggest that our GVD-CVAE is robust to the choice of the language decoder (rows 1,4-6), and achieves top grounding performance even when using a simple LSTM in the decoder or an UpDown LSTM with soft-attention only over grid features, demonstrating the effectiveness of our latent-variable modeling.

(3) What is the effect of the proposed training objective? We first train our GVD-CVAE with the vanilla CVAE loss, i.e., with $\lambda, \beta = 1$. Without any of our proposed modifications, this results in a very low grounding performance, as can be seen in the first row of Table 3. By adding the cross-entropy loss term that penalizes word predictions based on soft region context determined by the p-attention network (CE+ELBO), we are able to improve upon the soft-attention baseline of 22%. However, learning curves (included in the appendix) show that the KL loss term has vanished, suggesting that the model's posterior has collapsed to the prior and the approximate posterior alignment does not additionally take into account the word being grounded. Applying known solutions to KL vanishing, such as linearly annealing the β hyperparameter from 0 to 1 (CE+ELBO+β anneal), does not solve the problem. Instead, our proposed clipped linear annealing schedule leads to overall better grounding of 28.9% (KL term $\approx$ 0.06). Alternatively, after we determine a desirable value for the KL term, we can use the PI-Controller [55] anneal β, which we found to be less sensitive to changes in architecture and requires minimal calibration. Note that in this ablation we used a single LSTM language decoder and an LSTM in the inference model for faster experimentation.

4.3. Comparison with the State of the Art

As shown in Table 4, our GVD-CVAE improves weakly-supervised object grounding by 12% compared to the GVD method (21.4% to 33.7%) on the **F30k image dataset**. Thus, it sets the state-of-the-art *VOG* result, and reduces the gap with the fully-supervised GVD approach (41.4%). It also generates more grounded captions (higher $F1_{all}$ and $F1_{loc}$ scores) than all other methods, given the same features (from **G**VD). We even outperform methods using Scene Graphs [70] for grounding [73]. Note that the $F1_{all}$

Table 4. **Results on the Flickr30k Entities test set**. The performance of the fully-supervised GVD model (Sup.) is reported as an upper-bound to the weakly-supervised approaches. Types of model inputs during inference: region proposals extracted and encoded following **G**VD [74] or B**U**TD [2], or **S**cene-graphs [70]. † denotes models trained using auxiliary image-to-text matching models [33]. *RL* denotes models fine-tuned via Reinforcement Learning [49]. Note that results in the third block are obtained with different inputs, and thus they are not directly comparable to ours. We report average results for our GVD-CVAE after 5 random runs (standard deviations are included in the appendix).

		VOG	GVD					
			Captioning				Grounding	
	Feat	Acc	B@4	M	C	S	$F1_{all}$	$F1_{loc}$
GVD [74] (Sup.)	G	41.4	27.3	22.5	62.3	16.5	7.55	22.2
GVD [74]	G	21.4	26.9	22.1	60.1	16.1	3.88	11.7
GVD-Grd [74]	G	25.5	26.9	22.1	60.1	16.1	3.88	11.7
Cyclical [40]	G	-	26.6	22.3	60.9	16.3	4.85	13.4
DPA [37]	G	-	27.6	22.6	62.7	16.7	4.79	15.5
SCAN-RL [77] †	G	-	28.0	22.6	66.2	17.0	6.53	15.8
BUTD [2]	U	24.2	27.3	21.7	56.6	16.0	-	-
DPA [37]	U	-	27.2	22.3	60.8	16.3	5.45	15.3
Sub-GC [73]	S	-	28.5	22.3	61.9	16.4	5.98	16.5
SCAN-RL [77] †	U	-	**30.1**	22.6	**69.3**	16.8	**7.17**	17.5
GVD-CVAE	G	**33.7**	24.0	21.3	55.3	15.7	6.70	**19.2**
GVD-CVAE-RL	G	31.6	29.8	**23.1**	67.6	**17.2**	6.94	17.6

Table 5. **Results on the ActivityNet Entities validation set.** We report average results for our GVD-CVAE after 5 random runs.

	VOG	GVD					
		Captioning				Grounding	
	Acc	B@4	M	C	S	$F1_{all}$	$F1_{loc}$
GVD (Sup.) [74]	35.7	2.59	11.2	47.5	15.1	7.1	24.1
MIL-based							
NAFAE [56]	19.5	-	-	-	-	-	-
STVG [67]	21.1	-	-	-	-	-	-
SCL [64]	23.8	-	-	-	-	-	-
Captioning-based							
GVD [74]	14.9	2.28	10.9	45.6	**15.0**	3.7	12.7
GVD-Grd [74]	21.3	2.28	10.9	45.6	15.0	3.7	12.7
Cyclical [40]	-	**2.45**	**11.1**	**46.4**	14.8	4.7	15.8
GVD-CVAE	**23.9**	1.90	10.4	41.8	13.3	**5.8**	**21.7**

Table 6. **Results on the YouCook2 test set following the experimental setup of Shi et al. [56]**. GVD* denotes our implementation and training of the GVD model [74].

	Box accuracy (%)	
	macro	micro
Upper Bound	62.41	-
MIL-based methods		
DVSA-frm [29]	37.55	44.16
Zhou [75]	35.08	42.42
NAFAE [56]	40.71	46.33
STVG [67]	41.67	48.22
SCL [64]	**42.80**	**48.60**
Captioning-based methods		
GroundR [51]	19.94	-
GVD* [74]	37.40	44.15
GVD-CVAE (Ours)	38.85 ± 0.20	44.62 ± 0.09

scores obtained by both our CVAE-p (6.43%) and CVAE-q (6.70%) distributions outperform Cyclical [40] (4.85%) and DPA [37] (4.79%). This suggests that modeling alignments as latent variables works better than applying attention regularization techniques during training. Despite generating more grounded captions, our method has lower captioning metrics than SoTA methods, some of which apply reinforcement learning (RL). However, our language model can also be finetuned with a CIDEr-based SCST loss [49] (GVD-CVAE-RL), leading to competitive captioning metrics.

Results on the ANet video dataset (Table 5) show similar trends. Our GVD-CVAE yields better metrics when grounding ground-truth or generated sentences. It also outperforms video-tailored, video-to-text matching models, such as NAFAE [56]. Although powerful, these models cannot tackle the *WS-GVD* task. Since we evaluate only on the validation set, we did not select the model with best CIDEr score, or tune the learning rate based on it. This might have led to our slightly inferior captioning metrics compared to [40, 74], which used the validation set for selecting a model to be evaluated on the now closed test server.

We also compare our method to MIL-based grounding approaches in the YouCook2 test split in terms of Box Accuracy (additional metrics are reported in the appendix). As seen in Table 6, although our method outperformed all video-to-text-matching methods on the ANet video dataset, it is, for instance, lagging behind NAFAE [56] by around 2% on the YouCook2 dataset. A possible explanation is that, while in ANet grounding is evaluated on a single frame, in YouCook2 grounding predictions are evaluated in every frame. Therefore, MIL-based methods that model the consistency between the localized regions at each frame or model inter-object interactions perform better. We believe that extending our GVD-CVAE to model such relationships will improve these metrics, and we leave that to future work. Finally, we show qualitative image grounding results in Fig. 4.

Limitations. Similar to all other proposal-based approaches, our model's performance is limited by the quality of the region proposals. Also, our GVD-CVAE does not model the dependency between alignments for consecutive words. Finally, we applied the same framework for image and video object grounding to demonstrate its generality and effectiveness, without taking advantage of several inductive biases in the video domain, such as the visual similarity between grounded regions in consecutive frames.

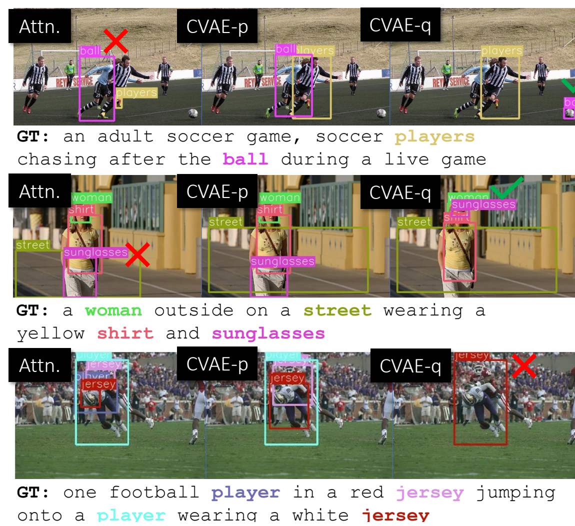

Figure 4. **Qualitative *WS-VOG* results on the F30k validation set**. For each ground-truth caption, we show grounding results obtained by (a) the soft-attention baseline, (b) our prior, and (c) our approximate posterior alignment distributions. We observe that knowing the words to be grounded improves grounding of small objects. Third row shows a failure case, in which our CVAE-q predicts the same bounding box for all groundable words.

5. Conclusion

In this paper, we proposed a novel grounded visual description CVAE. We showed how leveraging the latent alignment distributions of our model outperforms soft attention for grounding given ground-truth or generated sentences. We also demonstrated the generality and effectiveness of our model by evaluating it on both image and video datasets. Our novel approach yields competitive results in both grounding and grounded video description, while comparing against methods optimized for one of the two tasks.

Acknowledgements. The authors thank Benjamín Béjar Haro, Carolina Pacheco Oñate, Ambar Pal, Paris Giampouras, and the anonymous reviewers for their valuable comments. This research was supported by the IARPA DIVA program via contract number D17PC00345.

References

[1] Alexander A. Alemi, Ben Poole, Ian Fische, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a broken elbo. In *International Conference on Machine Learning*, volume 1, 2018.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[3] Jyoti Aneja, Harsh Agrawal, Dhruv Batra, and Alexander Schwing. Sequential latent spaces for modeling the intention during diverse image captioning. In *IEEE International Conference on Computer Vision*, pages 4260–4269, 2019.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[5] Apratim Bhattacharyya, Bernt Schiele, and Mario Fritz. Accurate and diverse sampling of sequences based on a 'best of many' sample objective. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[6] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *20th SIGNLL Conference on Computational Natural Language Learning*, 2016.

[7] Peter Carbonetto, Nando De Freitas, and Kobus Barnard. A statistical model for general contextual object recognition. *European Conference on Computer Vision*, 2004.

[8] Kan Chen, Jiyang Gao, and Ram Nevatia. Knowledge aided consistency for weakly supervised phrase grounding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4042–4050, 2018.

[9] Hong Min Chu, Chih Kuan Yeh, and Yu Chiang Frank Wang. Deep generative models for weakly-supervised multi-label classification. In *European Conference on Computer Vision*, pages 409–425. Springer Verlag, 2018.

[10] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *Neural Information Processing Systems*, volume 2015-January, 2015.

[11] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay DIvakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *IEEE International Conference on Computer Vision*, volume 2019-October, pages 2601–2610, 10 2019.

[12] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *IEEE International Conference on Computer Vision*, 2021.

[13] Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander M. Rush. Latent alignment and variational attention. In *Neural Information Processing Systems*, volume 2018-December, 2018.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Annual Meeting of the Association for Computational Linguistics*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[15] Adji B. Dieng, Yoon Kim, Alexander M. Rush, and David M. Blei. Avoiding latent variable collapse with generative skip models. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[16] Pelin Dogan, Leonid Sigal, and Markus Gross. Neural sequential phrase grounding (seqground). In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4175–4184, 2019.

[17] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, 2019.

[18] Anirudh Goyal, Alessandro Sordoni, Marc Alexandre Ct, Nan Rosemary Ke, and Yoshua Bengio. Z-forcing: Training stochastic recurrent networks. In *Neural Information Processing Systems*, volume 2017-December, 2017.

[19] Colin Graber and Alexander G. Schwing. Dynamic neural relational inference. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[20] Abhinav Gupta, Praveen Srinivasan, Jianbo Shi, and Larry S. Davis. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2012–2019, 2009.

[21] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, volume 12348 LNCS, 2020.

[22] Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. In *International Conference on Learning Representations*, 2019.

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[24] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pages 771–787, 2018.

[25] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[26] De An Huang, Shyamal Buch, Lucio Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. Finding 'it': Weakly-supervised reference-aware visual grounding in instructional videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[27] Pingping Huang, Jianhui Huang, Yuqing Guo, Min Qiao, and Yong Zhu. Multi-grained attention with object-level grounding for visual question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3595–3600, Florence, Italy, July 2019. Association for Computational Linguistics.

[28] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017.

[29] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:664–676, 4 2017.

[30] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. In *Foundations and Trends in Machine Learning*, 2019.

[31] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[32] Alon Lavie and Michael J Denkowski. The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23(2):105–115, 2009.

[33] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *European Conference on Computer Vision*, September 2018.

[34] Bohan Li, Junxian He, Graham Neubig, Taylor Berg-Kirkpatrick, and Yiming Yang. A surprisingly effective fix for deep latent variable modeling of text. In *Conference on Empirical Methods in Natural Language Processing*, 2020.

[35] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, 2020.

[36] Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. Attention correctness in neural image captioning. In *AAAI Conference on Artificial Intelligence*, 2017.

[37] Fenglin Liu, Xuancheng Ren, Xian Wu, Shen Ge, Wei Fan, Yuexian Zou, and Xu Sun. Prophet attention: Predicting attention with future attention. In *Neural Information Processing Systems*, 2020.

[38] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Neural Information Processing Systems*, pages 13–23, 2019.

[39] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.

[40] Chih-Yao Ma, Yannis Kalantidis, Ghassan AlRegib, Peter Vajda, Marcus Rohrbach, and Zsolt Kira. Learning to generate grounded visual captions without localization supervision. In *European Conference on Computer Vision*, 2020.

[41] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

[42] Arya D. McCarthy, Xian Li, Jiatao Gu, and Ning Dong. Addressing posterior collapse with mutual information for improved variational neural machine translation. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

[43] Artidoro Pagnoni, Kevin Liu, and Shangyan Li. Conditional variational autoencoder for neural machine translation, 12 2018.

[44] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[45] Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. Areas of attention for image captioning. In *IEEE International Conference on Computer Vision*, 2017.

[46] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123(1):74–93, 2017.

[47] Vignesh Ramanathan, Armand Joulin, Percy Liang, and Li Fei-Fei. Linking people in videos with "their" names using coreference resolution. In *European Conference on Computer Vision*, volume 8689 LNCS, 2014.

[48] Ali Razavi, Oriol Vinyals, Aron Van Den Oord, and Ben Poole. Preventing posterior collapse with δ-vaes. In *International Conference on Learning Representations*, 2019.

[49] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024, 2017.

[50] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018.

[51] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, 2016.

[52] Anna Rohrbach, Marcus Rohrbach, Siyu Tang, Seong Joon Oh, and Bernt Schiele. Generating descriptions with grounded and co-referenced people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4979–4989, 2017.

[53] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI Conference on Artificial Intelligence*, 2017.

[54] Shiv Shankar and Sunita Sarawagi. Posterior attention models for sequence to sequence learning. In *International Conference on Learning Representations*, 2019.

[55] Huajie Shao, Shuochao Yao, Dachun Sun, Aston Zhang, Shengzhong Liu, Dongxin Liu, Jun Wang, and Tarek Abdelzaher. Controlvae: Controllable variational autoencoder. In *International Conference on Machine Learning*, 2020.

[56] Jing Shi, Jia Xu, Boqing Gong, and Chenliang Xu. Not all frames are equal: Weakly-supervised video grounding with

contextual similarity and visual clustering losses. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[57] Mohit Shridhar and David Hsu. Interactive visual grounding of referring expressions for human-robot interaction. In *Proceedings of Robotics: Science and Systems*, 2018.

[58] Gunnar A. Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, João Carreira, Phil Blunsom, and Andrew Zisserman. Visual grounding in video for unsupervised word translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[59] Kihyuk Sohn, Xinchen Yan, and Honglak Lee. Learning structured output representation using deep conditional generative models. In *Neural Information Processing Systems*, 2015.

[60] Giorgos Tziafas and Hamidreza Kasaei. Few-shot visual grounding for natural human-robot interaction. In *2021 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, pages 50–55. IEEE, 2021.

[61] Aisha Urooj, Hilde Kuehne, Kevin Duarte, Chuang Gan, Niels Lobo, and Mubarak Shah. Found a reason for me? weakly-supervised grounded visual question answering using capsules. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8465–8474, June 2021.

[62] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 07-12-June-2015, pages 4566–4575. IEEE Computer Society, 10 2015.

[63] Liwei Wang, Alexander G. Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *Neural Information Processing Systems*, volume 2017-December, 2017.

[64] Wei Wang, Junyu Gao, and Changsheng Xu. Weakly-supervised video object grounding via stable context learning. In *ACM International Conference on Multimedia*, New York, NY, USA, 2021. Association for Computing Machinery.

[65] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2017-January, pages 5253–5262, 11 2017.

[66] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, 2015.

[67] Xun Yang, Xueliang liu, Meng Jian, Xinjian Gao, and Meng Wang. Weakly-supervised video object grounding by exploring spatio-temporal contexts. In *ACM International Conference on Multimedia*, New York, NY, USA, 2020. Association for Computing Machinery.

[68] Haonan Yu and Jeffrey Mark Siskind. Grounded language learning from video described with sentences. In *Annual Meeting of the Association for Computational Linguistics*, volume 1, 2013.

[69] Manzil Zaheer, Amr Ahmed, and Alexander J. Smola. Latent lstm allocation joint clustering and non-linear dynamic modeling of sequential data. In *International Conference on Machine Learning*, volume 8, 2017.

[70] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[71] Junchao Zhang and Yuxin Peng. Object-aware aggregation with bidirectional temporal graph for video captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[72] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Annual Meeting of the Association for Computational Linguistics*, 2017.

[73] Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. Comprehensive image captioning via scene graph decomposition. In *European Conference on Computer Vision*, 2020.

[74] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J. Corso, and Marcus Rohrbach. Grounded video description. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 6571–6580. IEEE Computer Society, 6 2019.

[75] Luowei Zhou, Nathan Louis, and Jason J. Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction. In *British Machine Vision Conference*, 2019.

[76] Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[77] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. More grounded image captioning by distilling image-text matching model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4776–4785, 8 2020.