Comprehensive Visual Grounding for Video Description

Wenhui Jiang[1], Yibo Cheng[1], Linxin Liu[1], Yuming Fang[1*], Yuxin Peng[2], Yang Liu[3]

[1]Jiangxi University of Finance and Economics, Nanchang, China
[2]Peking University, Beijing, China
[3]Sany Heavy Industry CO., LTD,China
jiang1st@bupt.cn, cyb592891032@gmail.com, linxinliu2010@gmail.com, fa0001ng@e.ntu.edu.sg, pengyuxin@pku.edu.cn, luy2655@sany.com.cn

Abstract

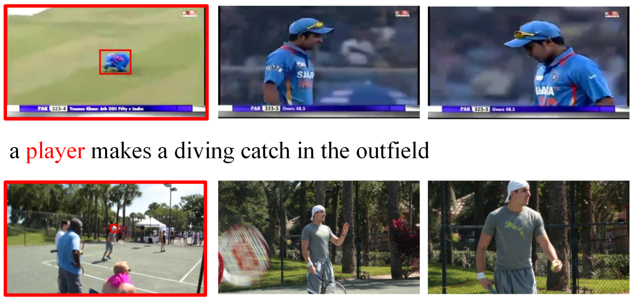
The grounding accuracy of existing video captioners is still behind the expectation. The majority of existing methods perform grounded video captioning on sparse entity annotations, whereas the captioning accuracy often suffers from degenerated object appearances on the annotated area such as motion blur and video defocus. Moreover, these methods seldom consider the complex interactions among entities. In this paper, we propose a comprehensive visual grounding network to improve video captioning, by explicitly linking the entities and actions to the visual clues across the video frames. Specifically, the network consists of spatial-temporal entity grounding and action grounding. The proposed entity grounding encourages the attention mechanism to focus on informative spatial areas across video frames, even if the entity is annotated in only one frame of a video. The action grounding dynamically associates the *verbs* to related subjects and the corresponding context, which keeps fine-grained spatial and temporal details for action prediction. Both entity grounding and action grounding are formulated as a unified task guided by a soft grounding supervision, which brings architecture simplification and improves training efficiency as well. We conduct extensive experiments on two challenging datasets, and demonstrate significant performance improvements of +2.3 CIDEr on ActivityNet-Entities and +2.2 CIDEr on MSR-VTT compared to state-of-the-arts.

Introduction

Video captioning aims to describe the visual content in the video using natural language sentences. It remains a challenging task as it requires a deep understanding of the objects and their interactions.

Existing methods for video captioning usually employ attention mechanisms, which are expected to ground correct visual regions for proper word generation. Although these models have achieved remarkable performance, previous researches (Zhou et al. 2019, 2020; Fei 2022) have shown that attention mechanisms are incapable of correctly associating generated words with meaningful visual regions, which makes the model less interpretable.

To address this problem, recent works (Liu et al. 2017; Jiang et al. 2022) have exploited region-phrase annotations in the training stage and designed diverse objective functions to guide the attention module to ground on appropriate visual areas. These methods achieve desirable improvements in still images. However, directly applying these grounding modules for video captioning is highly challenging due to the following reasons:



a player makes a diving catch in the outfield

several people are watching a man in a white hat on a tennis court

Figure 1: Examples of ActivityNet-Entities dataset. The frames with grounding annotations are marked in red. Only one bounding box per entity is labeled for each video.

1. Relevant visual regions corresponding to the object entities can span several frames. However, carefully labeling the bounding box of each entity frame by frame is labor-consuming. As is exemplified in Fig. 1, existing datasets provide only sparse annotations (Zhou et al. 2019), *i.e.*, annotating each entity with a bounding box in one frame of the video. Recent grounded video captioning models then perform spatial-level grounding within the annotated frame (Zhou et al. 2019; Wan, Jiang, and Fang 2022), ignoring the temporal dynamics of the entities across video frames.
2. Unlike image captioning that emphasizes the prediction of *nouns*, video descriptions are featured for the complex actions and interactions of objects. However, due to the lack of explicit visual annotations of *verbs*, action grounding remains challenging. Several methods (Ye et al. 2022; Zheng, Wang, and Tao 2020) associate *verbs* with global motion features, which may cause considerable spatial details missing.

To fully explore the spatial and temporal correlations

[*]Corresponding author: fa0001ng@e.ntu.edu.sg

among the video to achieve accurate grounded video captions, we propose a comprehensive visual grounding network. It performs spatial-temporal grounding on both entities and actions, aiming to predict accurate *nouns* and *verbs*.

For entity grounding (EG), our observation is that the annotated entities may suffer from deteriorated appearances in videos, such as motion blur, inappropriate viewpoint, *etc*. As a result, the labeled visual clues may not be informative enough to generate the target word. In contrast, recognizing the entities in adjacent frames can be easier (see Fig. 1 for illustrations). Therefore, we propose dynamic label propagation from the labeled frame to adjacent frames using "detection by tracking" strategy. The generated entity tracklet enables spatial-temporal entity grounding.

For action grounding, we are motivated by the syntax triplet ⟨subject, predicate, object⟩, where actions are always associated with subjects and objects. We therefore automatically generate grounding annotations of actions by referring to the union of the areas related to the subject, objects and corresponding context, and encourage the attention mechanism to ground on these areas.

To achieve video grounding, we further propose a soft grounding supervision (SGS) which encourages the attention mechanism grounding on informative spatial-temporal areas softly. The attention mechanism supervised by SGS enables the generated caption to have a reasonable meaning. More importantly, SGS unifies entity grounding and action grounding. The task unification not only simplifies the captioning architecture, but also improves training efficiency.

We evaluate our method on ActivityNet-Entities and MSR-VTT (Xu et al. 2016). Both quantitative and qualitative comparisons verify that our method significantly improves video captioning. Notably, our method achieves the CIDEr scores of 51.8 and 60.2 on ActivityNet-Entities and MSR-VTT, respectively, which are +2.3 and +2.2 higher than the best competitors.

In sum, the contributions of this work are threefold:

- Propose spatial-temporal entity grounding (EG) which dynamically focuses on informative spatial areas across video frames, albeit the entity is annotated in only one frame. EG strengthens the temporal context and improves visual *nouns* prediction.
- Propose action grounding that associates the actions to object-related spatial-temporal areas. To the best of our knowledge, there has not been any deep exploration of action grounding for video captioning.
- Propose a soft grounding supervision (SGS) that encourages the captioner grounding on informative spatial-temporal areas softly. SGS simplifies the grounding architecture and makes the captioning model interpretable.

Related Work

Video Captioning. Most recent video captioning methods employ an encoder-decoder framework. The encoder extracts video representations from a set of video frames, and the decoder generates the sentence word-by-word according to the video representation.

To improve vision-word alignment for precise word generation, Yao *et al.* (Yao et al. 2015) introduce a temporal attention mechanism into video captioning, which allows the decoder to automatically focus on the most relevant frames conditioned by the LSTM hidden state. Other representative methods (Yan et al. 2019; Chen and Jiang 2021; Tang et al. 2022) combine spatial and temporal attention. The spatial attention emphasizes important regions in each video frame for word generation. Meanwhile, the temporal attention is used to derive a subset of frames that is correlated to the video caption. More recently, researchers have exploited spatial and temporal attention simultaneously to build cross-modal associations. For example, LSRT (Li et al. 2022) builds spatial-temporal relationships between adjacent objects and proposes a coarse-to-fine decoder that attends to relevant objects spatially and temporally. SwinBERT (Lin et al. 2022) exploits Video Swin Transformer to encode spatial-temporal visual tokens and adopt a multimodal transformer that simultaneously attends to sparse visual tokens within the video to perform precise decoding. Although these works have significantly promoted video captioning, it is widely acknowledged that existing attention-based models are not correctly grounded.

Video Grounding. Video grounding aims to localize the starting and ending time of the target video segment that corresponds to the given text. Conventional methods (Liu et al. 2022a) firstly generate candidate proposals, then semantically match a given query text with each candidate through video-text matching. Yang *et al.* (Yang et al. 2022) further study spatio-temporal video grounding, which aims at localizing a spatio-temporal tube corresponding to the given text. Different from video grounding that leverages provided video captions for spatial-temporal localization, our work exploits video grounding as an intermediate step to improve captioning.

Recent efforts have been put into improving visual grounding for captioning. As a representative work, GLIPv2 (Zhang et al. 2022) takes visual grounding as model pre-training and takes image captioning as the downstream task. Different from GLIPv2, our work directly builds the connection between video grounding and video captioning. Other works focus on grounded captioning. Conventional methods (Jiang et al. 2022; Liu et al. 2017; Zhou et al. 2019) introduce an auxiliary task that builds correct correlations between object words and the corresponding image regions during the caption generation. For example, GVD (Zhou et al. 2019) explicitly links each *noun* phrase with the corresponding bounding box in one of the frames of a video. However, it only emphasizes the prediction of objects in the video, while ignoring the rich actions and events implied in the video. In addition, GVD only focuses on grounding on a single sampled frame, ignoring the temporal dynamics of the objects across frames. In contrast, SAAT (Zheng, Wang, and Tao 2020) and HMN (Ye et al. 2022) improve action prediction by associating actions with motion features. However, they only focus on temporal action correspondences, which disregard spatial details.

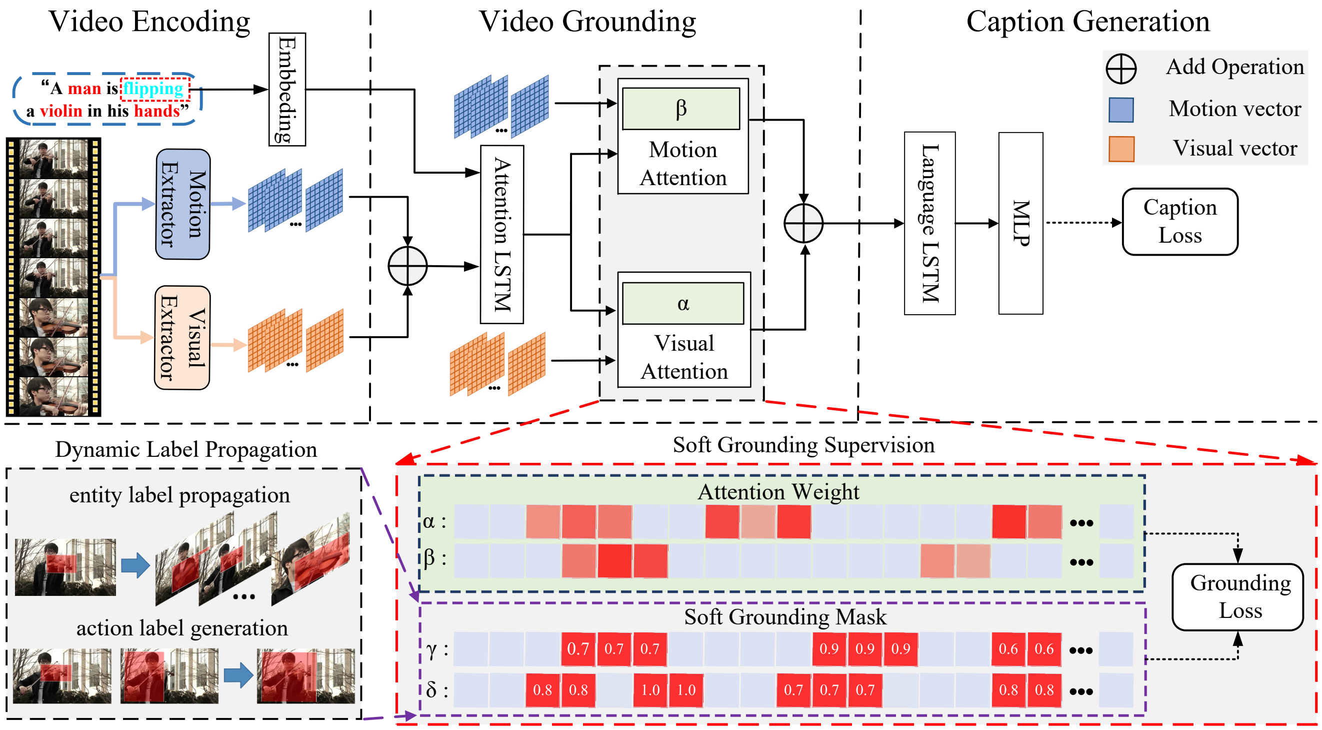


Figure 2: Overview of our comprehensive visual grounding for video captioning. Dynamic label propagation generates entity and action annotations for all video frames. Soft grounding supervision guides the attention mechanism dynamically attending to the relevant spatial-temporal regions of the input video for both entities and actions. Dynamic label propagation and soft grounding supervision are only employed in the training phase, and thus would not impact inference efficiency.

Proposed Method

Overview

The grounded video captioning model takes a sequence of raw video frames as inputs and outputs a caption Y. We denote the ground truth sentence with T words as $Y = (y^1, y^2, ..., y^T)$. Each *noun* y^t is associated with one bounding box annotation B^t that indicates its appearance in one of the video frames. A captioning model is learned to maximize the conditional probability $\mathbf{p}(Y|I;\theta)$ for each video, where θ denotes the model parameters.

The overall framework is shown in Fig. 2. We follow the conventional encoder-decoder pipeline, which consists of three modules, feature encoding, video grounding, and caption generation. The feature encoder extracts grid-level features to retain spatial-wise information of the video. The visual grounding module performs spatial-temporal entity grounding and action grounding, leading to an improved attention mechanism. We introduce dynamic label propagation to estimate target entities and action areas across the video frames. Both entity grounding and action grounding are formulated as a unified task guided by a soft grounding supervision and are learned to dynamically focus on the relevant spatial-temporal regions. The grounding module finally forms a high-level impression of the video content and feeds it for caption generation. We describe each module in detail as follows.

Video Encoding

The video encoder detects informative visual clues from the video. We employ a vision transformer as the video encoder to extract the visual feature $\mathbf{V}$ and motion feature $\mathbf{M}$. Specifically, we uniformly sample F frames from each video segment. These frames are then processed by the Video Swin Transformer to compute temporal-aware visual features. Each frame f produces N grid feature vectors $\mathbf{V}_f = [\mathbf{v}_f^1, \mathbf{v}_f^2, ...\mathbf{v}_f^N]$. To obtain the visual feature of the video, we concatenate the feature vectors from all frames, where the visual feature vector represented as $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_G]$, where G is the total number of visual grid vectors extracted from the entire video and $G = F \times N$. Meanwhile, we utilize Text4Vis model (Wu, Sun, and Ouyang 2023) to extract grid-level motion features $\mathbf{M}_f = [\mathbf{m}_f^1, \mathbf{m}_f^2, ..., \mathbf{m}_f^N]$ from the sampled frames. Similar to the visual features, we obtain the motion features as $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_H]$, where H denotes the total number of motion vectors extracted from the whole video.

Video Grounding

Attention Mechanism. The video attention learns to selectively attend to relevant visual areas for sentence generation. Following GVD (Zhou et al. 2019), we employ the widely used additive attention on visual features and motion features, respectively. Formally, at decoding time step t, an

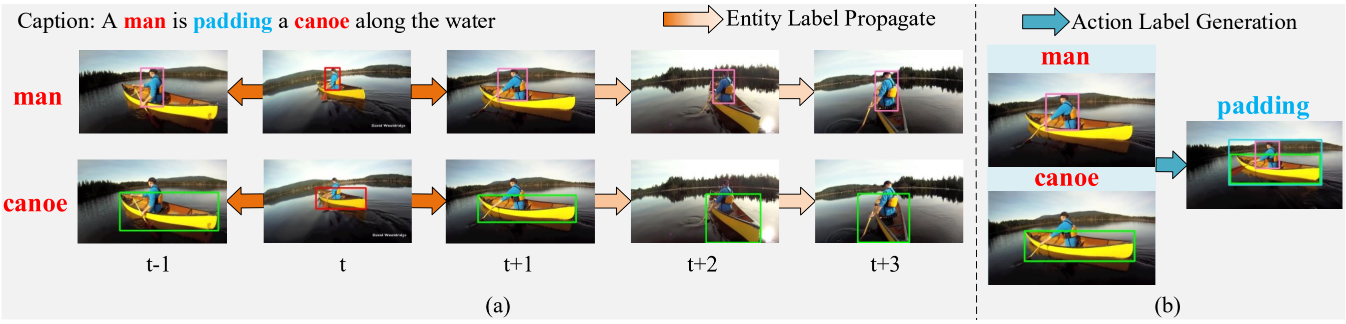


Figure 3: Illustration of dynamic label propagation. (a)Entity label propagation. (b)Action label generation.

attention LSTM first takes the visual feature, motion feature and word embedding of input word y^{t-1} as inputs, and outputs a hidden state $\mathbf{h}_1^t$:

$$\mathbf{h}_1^t = \mathrm{LSTM}_1([\bar{\mathbf{V}} + \bar{\mathbf{M}}; \mathbf{W}_e y^{t-1}], \mathbf{h}_1^{t-1}) \tag{1}$$

where $\bar{\mathbf{V}} = \frac{1}{G}\sum_{i=1}^{G} \mathbf{v}_i; \bar{\mathbf{M}} = \frac{1}{H}\sum_{i=1}^{H} \mathbf{m}_i$, $\mathbf{W}_e$ denotes the word embedding matrix. Then, conditioned on the hidden state h_1^t, we can calculate the attention distribution α^t and β^t for $\mathbf{V}$ and $\mathbf{M}$ as follows:

$$\begin{aligned} \alpha^t &= \mathrm{softmax}(\mathbf{W}_a^{\alpha}[\tanh(\mathbf{W}_k^{\alpha}\mathbf{V} + \mathbf{W}_q^{\alpha}\mathbf{h}_1^t)]) \\ \beta^t &= \mathrm{softmax}(\mathbf{W}_a^{\beta}[\tanh(\mathbf{W}_k^{\beta}\mathbf{M} + \mathbf{W}_q^{\beta}\mathbf{h}_1^t)]) \end{aligned} \tag{2}$$

where $\mathbf{W}_a$, $\mathbf{W}_k$ and $\mathbf{W}_q$ are the embedding matrices, $\alpha^t = [\alpha_1^t, \alpha_2^t, ..., \alpha_G^t]$ and $\beta^t = [\beta_1^t, \beta_2^t, ..., \beta_H^t]$ denote the attention weights for $\mathbf{V}$ and $\mathbf{M}$, respectively. For simplicity, we drop the superscript t in the rest of the paper unless explicitly mentioned.

Dynamic Label Propagation. As α and β are learned as latent variables without explicit supervision, the attention models are criticized for the "deviated focus" problem (Fei 2022; Jiang et al. 2022). The most straightforward way to overcome this issue is to introduce grounding supervision. However, existing video captioning datasets annotate only one bounding box for each entity throughout the video sequence, which provides insufficient visual clues in case of small object scale and inappropriate viewpoint. Moreover, no action annotation is provided. To fully leverage the spatial and temporal correlations among the video, we propose dynamic label propagation (DLP).

For entities, the DLP generates pseudo box annotations in adjacent video frames using "detection by tracking" strategy. Specifically, as shown in Fig. 3(a), for each entity with a labeled bounding box B, we leverage ToMP (Mayer et al. 2022), a high-performing object tracker, to generate a tracklet over the entire video. ToMP eventually outputs a pseudo annotation B_f for each unlabeled frame. Each B_f is also associated with a score s_f ranging from 0 to 1, indicating the confidence of the identified box. As ToMP may generate boxes with wrong locations and false positives, we apply confidence-based thresholding to further reduce potentially wrong pseudo boxes, *i.e.*, only pseudo annotations with confidence scores higher than s_{th} are maintained.

A critical issue for action grounding is the lack of visual annotations for action words. We observe that action words are always associated with entities (*i.e.*, subjects and objects). We therefore automatically generate grounding annotations for actions by referring to the union of the areas related to the entities of the action. The procedure is exemplified in Fig.3(b). Formally, for video frame f, given K associated entities and the corresponding boxes $\{B_f^i\}_{i=1}^K$, we generate the tightest bounding rectangle that covers these boxes as the annotation for the action, denoted by B'_f. We also generate a confidence score for B'_f by aggregating the scores from action-related entities:

$$s'_f = \min_i \{s_f^i\} \tag{3}$$

The generated action annotations allow us to build entity grounding and action grounding as a unified model easily.

Soft Grounding Supervision. Subsequently, we propose a soft grounding supervision (SGS) that encourages the attention mechanism grounding on informative spatial-temporal areas for both entities and actions. We take an example of grounding an entity word for illustration. The motivation behind the supervision is that α and β should be more concentrated on the annotated spatial-temporal areas with higher s_f. To that end, we construct a sequence of heatmaps $\mathbf{\Gamma} = [\mathbf{\Gamma}_1, \mathbf{\Gamma}_2, ..., \mathbf{\Gamma}_F]$ on visual grid features, where $\mathbf{\Gamma}_f = [\gamma_f^1, \gamma_f^2, ..., \gamma_f^N]$ has the same spatial resolution as $\mathbf{V}_f$. We encourage the attention model to focus on annotated areas by setting $\mathbf{\Gamma}_f$ with soft scores[1]:

$$\gamma_f^i = \begin{cases} s_f & i \in B_f \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

In order to facilitate calculation, we flatten heatmaps $\mathbf{\Gamma}$ into a vector $\gamma = [\gamma_1, \gamma_2, ..., \gamma_G]$. Similarly, we construct heatmaps on motion features and flatten them into a vector δ. The per-word grounding loss function is defined as follows:

$$\mathcal{L}_{vg} = -\log(\sum_{j=1}^{G} \gamma_j \alpha_j) - \log(\sum_{j=1}^{H} \delta_j \beta_j) \tag{5}$$

[1] We set the confidence score of the labeled box B to 1.

Eq.(5) encourages most grid features located inside annotated boxes to output high attention scores. Here γ and δ serve as soft voters. B_f with higher s_f is more likely to be concentrated since a larger γ_j enforces a larger α_j.

The same loss is applied to action words. The only difference is to use B'_f instead of B_f for heatmaps generation. The final loss on $\mathcal{L}_{vg}$ is the average of losses on all visually groundable words. In contrast to previous methods (Ye et al. 2022; Zheng, Wang, and Tao 2020) that refine entity prediction and action prediction with different designs, our work unifies entity and action grounding, which simplifies the captioning architecture and improves training efficiency.

Caption Generation

For caption generation, we apply the widely used attention-enhanced LSTM decoder. The video grounding module finally forms a high-level impression of the visual content by accumulating $\mathbf{V}$ and $\mathbf{M}$ with the attention weights:

$$\mathbf{q}^t = \sum_{j=1}^{G} \alpha_j \mathbf{v}_j + \sum_{j=1}^{H} \beta_j \mathbf{m}_j \tag{6}$$

where $\mathbf{q}^t$ corresponds specifically to individual words being generated, and is fed into a language LSTM to predict the next word y_t^*:

$$\mathbf{h}_2^t = \mathrm{LSTM}_2([\mathbf{q}^t; \mathbf{h}_A^t], \mathbf{h}_L^{t-1}) \tag{7}$$

$$y_t^* \sim \mathbf{p}^t = \mathrm{softmax}(\mathbf{W}_s \mathbf{h}_2^t) \tag{8}$$

where $\mathbf{W}_s$ is the embedding matrix, $\mathbf{p}^t$ denotes the output probability distribution of the decoder, and the generated word y_t^* is sampled from $\mathbf{p}^t$.

Training Objectives

We formulate the grounded video captioning as a joint optimization over the language and grounding tasks. The overall objective function is defined as follows:

$$\mathcal{L} = \mathcal{L}_{cap} + \lambda \mathcal{L}_{vg} \tag{9}$$

where $\mathcal{L}_{cap}$ denotes the caption generation loss, which compares the output sentence with the ground truth. Specifically, we employ the cross-entropy loss as follows:

$$\mathcal{L}_{cap} = -\sum_{t=1}^{T} \log(\mathbf{p}^t(y^t | y^{1:t-1})) \tag{10}$$

$\mathcal{L}_{vg}$ corresponds to the soft grounding supervision. λ is used to balance the two types of losses. $\mathcal{L}_{vg}$ serves as a word-region alignment regularization, which assists the captioning model in attending to informative regions.

Experiments

Experimental Setups

Datasets. We conduct our experiments on ActivityNet-Entities and MSR-VTT. The ActivityNet-Entities not only contains the video caption annotation of the video but also provides the box annotation of the *noun* phrase in the caption. The dataset contains 15,000 videos, including 52,000 video segments, and 1 caption annotation for each video segment. The dataset provides a total of 158,000 valid box annotations of 432 classes. The MSR-VTT contains 10,000 video clips from YouTube. There are 20 human descriptions for each video clip. The dataset contains 6,573 samples for training, 497 samples for validation, and 2,990 for testing.

EG	DLP	AG	SGS	B@1	B@4	S	M	C
-	-	-	-	24.6	3.1	16.0	11.7	51.9
✓	-	-	-	25.0	3.2	16.1	11.8	52.6
✓	✓	-	-	24.8	3.1	16.0	11.7	53.0
-	-	✓	-	25.0	3.2	16.0	11.8	52.8
✓	-	✓	✓	**25.1**	**3.3**	16.3	11.8	53.1
✓	✓	✓	-	25.0	3.2	16.2	11.8	53.5
✓	✓	✓	✓	**25.1**	**3.3**	**16.6**	**11.9**	**54.5**

Table 1: Ablation studies on ActivityNet-Entities val set. B@N, M, S, and C stand for BLEU@N, METEOR, SPICE, and CIDEr, respectively. The symbol "✓" indicates the inclusion of the following component. Bold for the best.

Evaluation Metrics. Following the standard video captioning evaluation protocol, we use 5 common captioning metrics to evaluate the captioning quality, *i.e.*, CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), BLEU (Papineni et al. 2002), METEOR (Denkowski and Lavie 2014), ROUGE-L (Lin 2004) and SPICE (Anderson et al. 2016).

Implementation Details. For ActivityNet-Entities, we uniformly sample 10 frames for each video segment. For MSR-VTT, 32 video frames are sampled from the video clip. We employ Video Swin Transformer (Liu et al. 2022c) pre-trained on ImageNet (Deng et al. 2009) to extract visual features. Besides, we use Text4Vis (Wu, Sun, and Ouyang 2023) model pre-trained on Kinetics-400 (Carreira and Zisserman 2017) to extract motion features.

The same model hyperparameters and data preprocessing step as GVD are adopted. The word embedding size is set to 512. Empirically, λ is set to 0.1. During training, we optimize the model with Adam for 25 epochs. The learning rate is initialized to be 5e-4 and decayed by a factor of 0.8 every three epochs.

Ablation Studies

To quantify the impact of different components for video captioning, we conduct ablation studies on the ActivityNet-Entities val set.

Effectiveness of Grounding Modules. To show the effectiveness of the grounding modules, we compare different variants of the proposed model. For clarity, we disable the soft grounding supervision. We start from the baseline model without any grounding modules, then we gradually incorporate entity grounding (EG) and action grounding

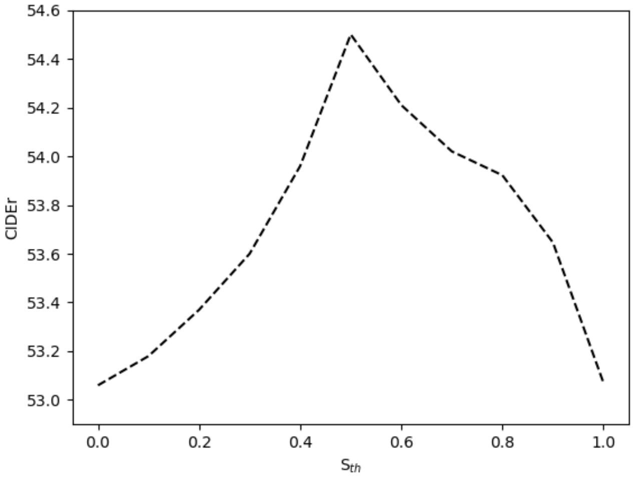


Figure 4: The impact of s_{th} on grounded captioning performance. A large s_{th} indicates high precision but low recall of the generated annotations. Experiments are conducted on ActivityNet-Entities val set.

(AG) to examine the effectiveness. As shown in Table 1, performing entity grounding on the labeled boxes solidly promotes the baseline. With dynamic label propagation (DLP), the entity grounding further improves the captioning performance overall metrics substantially, which suggests that the spatial-temporal entity grounding can exploit more temporal contexts to better solve video captioning. We also notice that action grounding alone improves the video captioning baseline significantly, which verifies that action grounding is critical for predicting *verbs*. Finally, when we integrate both grounding modules together, we obtain a CIDEr of 53.5, which outperforms the baseline by 1.6, thus demonstrating the superiority of the proposed grounding model.

The Impact of Label Propagation. We further investigate the importance of the proposed dynamic label propagation (DLP). Towards this goal, we adjust s_{th} to see how label propagation impacts video grounding. We notice that entity label propagation with varying s_{th} achieves consistently better captioning performance compared with the baseline. Specifically, when s_{th} varies from 0 to 0.5, the performance monotonically increases. The reason is that a high s_{th} ensures the accuracy of generated pseudo annotations, while the defective annotations may mislead the attention module conversely. Then the captioning performance reaches the peak when s_{th} reaches 0.5. Finally, the performance drops as s_{th} continues increasing. This is probably due to informative entities in adjacent frames being filtered as s_{th} goes higher. In an extreme case, setting s_{th} to 1.0 is equivalent to disabling entity label propagation. In the following experiments, we set s_{th} to 0.5 if not otherwise specified.

Effectiveness of Soft Grounding Supervision. To explore the effect of soft grounding supervision (SGS), we compare it with baseline supervision which treats all generated annotations as equally important. As reported in Table 1, soft grounding further promotes the performance of video captioning substantially. Specifically, we achieved a performance of 54.5 for CIDEr score, which is +2.6 higher than the baseline. Consistent performance boosts are observed on all other metrics. The reason is that soft grounding reduces the noise brought by generated pseudo annotations.

Method	B@1	B@4	S	M	C
Mask-TF	22.9	2.41	13.7	10.6	46.1
BiLSTM+TA	22.8	2.17	11.8	10.2	42.2
Cyclical	23.4	2.43	14.3	10.8	46.6
GVD	23.6	2.35	14.7	11.0	45.5
KNN-HAST	-	2.61	15.1	11.3	48.5
IAS	24.2	2.76	-	11.3	49.5
SwinBERT*	21.4	1.97	16.2	10.5	39.3
VIOLETv2*	21.4	1.83	15.2	10.7	38.5
Ours	**24.8**	**3.00**	**16.3**	**11.8**	**51.8**

Table 2: Comparisons of the state-of-the-art methods on ActivityNet-Entities test set. * denotes our re-implementation.

Comparison with State-of-the-art

Results on ActivityNet-Entities. We compare our model with several recent methods, including Mask-TF (Zhou et al. 2018), BiLSTM+TA (Zhou et al. 2018), Cyclical (Ma et al. 2020), GVD (Zhou et al. 2019), KNN-HAST (Shen et al. 2020), IAS (Wan, Jiang, and Fang 2022), SwinBERT (Lin et al. 2022) and VIOLETv2 (Fu et al. 2023).

Table 2 shows the detailed comparisons. It is clear that our model consistently exhibits better performance than the other competitors in terms of all metrics by a large margin. To be specific, our method achieves the performance of 51.8 for CIDEr score, which is +2.3 higher than that of IAS, the best-performing grounded video captioning model. We observe similar improvements in other evaluation metrics. It is worth noticing that our method achieves more significant improvements over recent video captioning models, *e.g.*, SwinBERT (+9.9 on CIDEr) and VIOLETv2 (+13.3 on CIDEr). The reason is that SwinBERT and VIOLETv2 only employ Video Swin Transformer as the video encoder, which may lack explicit motion information that is valuable for predicting the complex actions in ActivityNet-Entities.

Results on MSR-VTT. We further conduct analysis on the challenging MSR-VTT dataset. As MSR-VTT does not include any grounding annotations, we employ GLIPv2 (Zhang et al. 2022), a prevailing open-vocabulary object detector, to generate the most confidence bounding box as the seed label for each entity.

Table 3 summarizes the performance of our method and existing state-of-the-art methods, including MARN (Pei et al. 2019), OA-BTG (Zhang and Peng 2019), SAAT (Zheng, Wang, and Tao 2020), ORG-TRL (Zhang et al. 2020), UniVL (Luo et al. 2020), LSRT (Li et al. 2022), SwinBERT (Lin et al. 2022), HMN (Ye et al. 2022), BME-WCO (Liu et al. 2022b), MELTR (Ko et al.

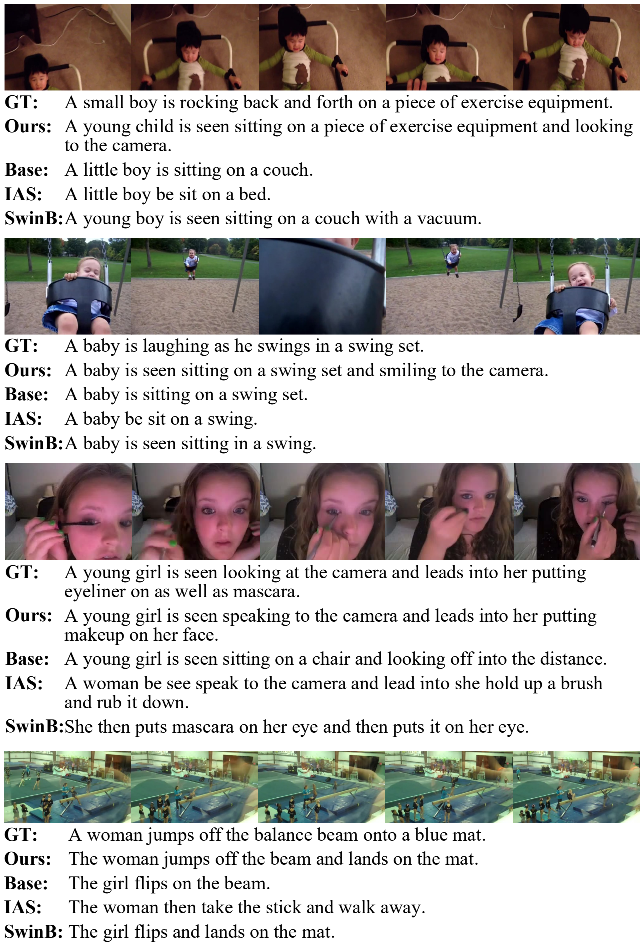
Method	B@1	B@4	M	R	C
MARN	-	40.4	28.1	60.7	47.1
OA-BTG	-	41.4	28.2	-	46.9
SAAT	79.6	39.9	27.7	61.2	51.0
ORG-TRL	-	43.6	28.8	62.1	50.9
UniVL	-	41.8	28.9	60.8	50.0
LSRT	-	42.6	28.3	61.0	49.5
SwinBERT	83.1	41.9	29.9	62.1	53.8
HMN	81.3	43.5	29.0	62.7	51.5
BME-WCO	-	40.6	28.1	61.2	53.4
MELTR	-	44.2	29.3	62.4	52.8
MAN	-	41.3	28.0	61.4	49.8
RSFD	-	43.4	29.3	62.3	53.1
VL-Prompt	-	43.2	30.1	62.7	55.3
VIOLETv2	-	-	-	-	58.0
Ours	**84.8**	**46.5**	**31.2**	**64.6**	**60.2**

Table 3: Comparisons of the state-of-the-art methods on MSR-VTT test set.

2023), MAN (Jing et al. 2023), RSFD (Zhong et al. 2023), VL-Prompt (Yan et al. 2023) and VIOLETv2 (Fu et al. 2023). As it can be observed, our method reaches 60.1 in terms of CIDEr and 46.5 in terms of BLEU@4, surpassing all other approaches significantly. Notably, our method significantly outperforms HMN (+8.7 on CIDEr score), which is designed to enhance entity and predicate generation. Our model also advances UniVL, MELTR and VIOLETv2 across all metrics, which leverage large-scale vision and language pretraining. These results solidly verify the superiority of the proposed method.

Qualitative Results

Fig. 5 showcases qualitative examples of the captions generated by SwinBERT, IAS and our method. The advantages of our method can be divided into two primary categories. Firstly, our method can recognize the entities more accurately. For example, in the first example, the "exercise equipment" is wrongly recognized as "a couch" in SwinBERT and "a bed" in IAS, while our method depicts the entity more precisely. In addition, our method provides fine-grained descriptions of the entities, benefiting from the rich temporal information of the entities across frames. For example, in the second example, our method enriches "a baby" with the attribute "smiling to the camera". Secondly, our method provides richer and more accurate content about the actions. As illustrated in the third example, our method predicts "put makeup", which is more informative than "hold up a brush","looking off into the distance" provided by other methods. The reason is that our method is learned to focus on informative regions related to the subjects to predict actions. In the last example, our method provides comprehensive description of the actions as "jumps off the beam" and "lands on the mat". More examples are presented in the supplementary.



GT: A small boy is rocking back and forth on a piece of exercise equipment.
Ours: A young child is seen sitting on a piece of exercise equipment and looking to the camera.
Base: A little boy is sitting on a couch.
IAS: A little boy be sit on a bed.
SwinB: A young boy is seen sitting on a couch with a vacuum.

GT: A baby is laughing as he swings in a swing set.
Ours: A baby is seen sitting on a swing set and smiling to the camera.
Base: A baby is sitting on a swing set.
IAS: A baby be sit on a swing.
SwinB: A baby is seen sitting in a swing.

GT: A young girl is seen looking at the camera and leads into her putting eyeliner on as well as mascara.
Ours: A young girl is seen speaking to the camera and leads into her putting makeup on her face.
Base: A young girl is seen sitting on a chair and looking off into the distance.
IAS: A woman be see speak to the camera and lead into she hold up a brush and rub it down.
SwinB: She then puts mascara on her eye and then puts it on her eye.

GT: A woman jumps off the balance beam onto a blue mat.
Ours: The woman jumps off the beam and lands on the mat.
Base: The girl flips on the beam.
IAS: The woman then take the stick and walk away.
SwinB: The girl flips and lands on the mat.

Figure 5: Examples of captions generated by our method and several state-of-the-art methods, as well as the corresponding ground-truths.

Conclusion

In this work, we aim to enhance the accuracy of grounded video captioning by introducing a comprehensive visual grounding network. This network comprises spatial-temporal entity grounding and action grounding. The entity grounding is responsible for directing attention to relevant spatial areas over the entire video, leveraging entity labels in only one frame of the video. In this meanwhile, the action grounding dynamically associates actions with relevant subjects and their respective contexts. This association allows the model to capture fine-grained spatial and temporal details necessary for accurate action prediction. Both entity grounding and action grounding are guided by a soft grounding supervision (SGS). SGS encourages the attention mechanism grounding on informative spatial-temporal areas softly. More importantly, SGS unifies entity grounding and action grounding, which simplifies the captioning architecture and improves training efficiency. Extensive experiments have demonstrated the superiority of our method compared with the state-of-the-arts.

Acknowledgements

This work was supported by the National Key Research and Development Program of China under Grant 2020AAA0109301.

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, 382–398.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Chen, S.; and Jiang, Y.-G. 2021. Motion guided region message passing for video captioning. In *IEEE International Conference on Computer Vision*, 1543–1552.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Denkowski, M.; and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Statistical Machine Translation*, 376–380.

Fei, Z. 2022. Attention-aligned transformer for image captioning. In *AAAI Conference on Artificial Intelligence*, 1, 607–615.

Fu, T.-J.; Li, L.; Gan, Z.; Lin, K.; Wang, W. Y.; Wang, L.; and Liu, Z. 2023. An empirical study of end-to-end video-language transformers with masked visual modeling. In *IEEE Conference on Computer Vision and Pattern Recognition*, 22898–22909.

Jiang, W.; Zhu, M.; Fang, Y.; Shi, G.; Zhao, X.; and Liu, Y. 2022. Visual cluster grounding for image captioning. *IEEE Transactions on Image Processing*, 31: 3920–3934.

Jing, S.; Zhang, H.; Zeng, P.; Gao, L.; Song, J.; and Shen, H. T. 2023. Memory-based Augmentation Network for Video Captioning. *IEEE Transactions on Multimedia*, 1–13.

Ko, D.; Choi, J.; Choi, H. K.; On, K.-W.; Roh, B.; and Kim, H. J. 2023. MELTR: Meta Loss Transformer for Learning to Fine-tune Video Foundation Models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 20105–20115.

Li, L.; Gao, X.; Deng, J.; Tu, Y.; Zha, Z.-J.; and Huang, Q. 2022. Long short-term relation transformer with global gating for video captioning. *IEEE Transactions on Image Processing*, 31: 2726–2738.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Lin, K.; Li, L.; Lin, C.-C.; Ahmed, F.; Gan, Z.; Liu, Z.; Lu, Y.; and Wang, L. 2022. Swinbert: End-to-end transformers with sparse attention for video captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 17949–17958.

Liu, C.; Mao, J.; Sha, F.; and Yuille, A. 2017. Attention correctness in neural image captioning. In *AAAI Conference on Artificial Intelligence*, 1.

Liu, D.; Qu, X.; Wang, Y.; Di, X.; Zou, K.; Cheng, Y.; Xu, Z.; and Zhou, P. 2022a. Unsupervised temporal video grounding with deep semantic clustering. In *AAAI Conference on Artificial Intelligence*, 2, 1683–1691.

Liu, S.; Li, A.; Wang, J.; and Wang, Y. 2022b. Bidirectional maximum entropy training with word co-occurrence for video captioning. *IEEE Transactions on Multimedia*, 1–1.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022c. Video swin transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3202–3211.

Luo, H.; Ji, L.; Shi, B.; Huang, H.; Duan, N.; Li, T.; Li, J.; Bharti, T.; and Zhou, M. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

Ma, C.-Y.; Kalantidis, Y.; AlRegib, G.; Vajda, P.; Rohrbach, M.; and Kira, Z. 2020. Learning to generate grounded visual captions without localization supervision. In *European Conference on Computer Vision*, 353–370.

Mayer, C.; Danelljan, M.; Bhat, G.; Paul, M.; Paudel, D. P.; Yu, F.; and Van Gool, L. 2022. Transforming model prediction for tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 8731–8740.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Pei, W.; Zhang, J.; Wang, X.; Ke, L.; Shen, X.; and Tai, Y.-W. 2019. Memory-attended recurrent network for video captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 8347–8356.

Shen, K.; Wu, L.; Xu, F.; Tang, S.; Xiao, J.; and Zhuang, Y. 2020. Hierarchical Attention Based Spatial-Temporal Graph-to-Sequence Learning for Grounded Video Description. In *International Joint Conference on Artificial Intelligence*, 941–947.

Tang, M.; Wang, Z.; Zeng, Z.; Li, X.; and Zhou, L. 2022. Stay in Grid: Improving Video Captioning via Fully Grid-level Representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33: 3319–3332.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4566–4575.

Wan, B.; Jiang, W.; and Fang, Y. 2022. Informative attention supervision for grounded video description. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1955–1959.

Wu, W.; Sun, Z.; and Ouyang, W. 2023. Revisiting classifier: Transferring vision-language models for video recognition. In *AAAI Conference on Artificial Intelligence*, 3, 2847–2855.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5288–5296.

Yan, C.; Tu, Y.; Wang, X.; Zhang, Y.; Hao, X.; Zhang, Y.; and Dai, Q. 2019. STAT: Spatial-temporal attention mechanism for video captioning. *IEEE Transactions on Multimedia*, 22(1): 229–241.

Yan, L.; Han, C.; Xu, Z.; Liu, D.; and Wang, Q. 2023. Prompt Learns Prompt: Exploring Knowledge-Aware Generative Prompt Collaboration for Video Captioning. In *International Joint Conference on Artificial Intelligence*.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2022. Tubedetr: Spatio-temporal video grounding with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, 16442–16453.

Yao, L.; Torabi, A.; Cho, K.; Ballas, N.; Pal, C.; Larochelle, H.; and Courville, A. 2015. Describing videos by exploiting temporal structure. In *IEEE International Conference on Computer Vision*, 4507–4515.

Ye, H.; Li, G.; Qi, Y.; Wang, S.; Huang, Q.; and Yang, M.-H. 2022. Hierarchical modular network for video captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 17939–17948.

Zhang, H.; Zhang, P.; Hu, X.; Chen, Y.-C.; Li, L. H.; Dai, X.; Wang, L.; Yuan, L.; Hwang, J.-N.; and Gao, J. 2022. GLIPv2: Unifying Localization and Vision-Language Understanding. In *Advances in Neural Information Processing Systems*.

Zhang, J.; and Peng, Y. 2019. Object-aware aggregation with bidirectional temporal graph for video captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 8327–8336.

Zhang, Z.; Shi, Y.; Yuan, C.; Li, B.; Wang, P.; Hu, W.; and Zha, Z.-J. 2020. Object relational graph with teacher-recommended learning for video captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 13278–13288.

Zheng, Q.; Wang, C.; and Tao, D. 2020. Syntax-aware action targeting for video captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 13096–13105.

Zhong, X.; Li, Z.; Chen, S.; Jiang, K.; Chen, C.; and Ye, M. 2023. Refined semantic enhancement towards frequency diffusion for video captioning. In *AAAI Conference on Artificial Intelligence*, 3724–3732.

Zhou, L.; Kalantidis, Y.; Chen, X.; Corso, J. J.; and Rohrbach, M. 2019. Grounded video description. In *IEEE Conference on Computer Vision and Pattern Recognition*, 6578–6587.

Zhou, L.; Zhou, Y.; Corso, J. J.; Socher, R.; and Xiong, C. 2018. End-to-end dense video captioning with masked transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 8739–8748.

Zhou, Y.; Wang, M.; Liu, D.; Hu, Z.; and Zhang, H. 2020. More grounded image captioning by distilling image-text matching model. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4777–4786.