DIBS: Enhancing Dense Video Captioning with Unlabeled Videos via Pseudo Boundary Enrichment and Online Refinement

Hao Wu[1,2] ♣, Huabin Liu[2,3] ♣, Yu Qiao[2], Xiao Sun[2] ✉

[1] University of Science and Technology of China [2] Shanghai Artificial Intelligence Laboratory

[3] Shanghai Jiao Tong University

wuhao@mail.ustc.edu.cn, huabinliu@sjtu.edu.cn, {qiaoyu, sunxiao}@pjlab.org.cn

1. Selection Process for Pretraining Subset: Focused on Cooking Videos

For our pretraining process, we carefully selected a subset of HowTo100M [2] videos, leveraging available metadata to curate a tailored dataset. Focusing on cooking-related content, we filtered videos based on the ”Recipes” category_2 tag, ensuring relevance to recipe-based tasks. Within this subset, we further refined our selection by prioritizing videos within the top 30 rankings. This selection strategy aimed to capture cooking-related content, given its sequential nature and specificity in depicting step-by-step instructions. Cooking videos often present a structured flow, mirroring the requirements for dense video captioning tasks. Their explicit actions and clear events make them ideal for training models to capture and describe such events accurately. Additionally, these instructional videos offer rich visual and textual information, providing diverse examples pivotal for robust model training and the generation of precise captions for subsequent tasks.

2. Prompt Engineering

To streamline the process of deriving step-by-step event sequences from video subtitles, we utilize LLMs like LLAMA2 [3] and ChatGPT to generate prompts. Initially, our requests for prompts, such as '*Task: write a prompt to send to LLM to extract the ordered steps of a recipe video subtitle. Example steps should resemble ”coat the chicken pieces in the flour”*,' resulted in step-by-step event captions that did not align with the concise, action-oriented format found in YouCook2 [4] captions.

To enhance this extraction process further, we manually refine the prompts to mirror the succinct nature of YouCook2 captions. Specifically, we emphasize extra regularization highlighting single-sentence clarity, exclusion of events unrelated to video actions, and emphasis on the sequential process. Additionally, after the instruction prompt, we append the original subtitle text followed by ”\nSteps:” to extract the step-by-step events without extra irrelevant words. The complete input for LLM should resemble ”*Prompt* $[original\ subtitle]$ *\nSteps:*”. This fine-tuning aims to closely align the extracted events with the concise and action-focused nature of YouCook2 captions, potentially shedding light on why pretraining shows better performance in YouCook2 [4] compared to ActivityNet [1]. Some potential prompt candidates are detailed in Table 1.

3. Additional ablation studies

Effects of Pretraining Data Variation Our comparative experiments in Table 2 exploring varying pretraining data proportions on YouCook2 and ActivityNet reveal a notable trend. Increasing the pretraining data volume significantly boosts model performance on both event localization and caption generation. Strikingly, we witness a substantial performance jump on YouCook2 compared to ActivityNet. This suggests the instructional cooking focus within YouCook2 derives greater benefit from more pretraining data. The cooking domain aligns closely with the pretraining data, enabling more efficient learning. In contrast, ActivityNet's broader activity spectrum may need more diverse pretraining data to show similar improvements. This highlights the pivotal role of dataset characteristics and domain relevance in pretraining efficacy for dense video captioning.

Pseudo Boundary Generation In our comparative study across different pseudo boundary generation schemes on the YouCook2 dataset, we evaluate the performance of models using uniform boundaries, Drop-DTW boundaries, and our proposed generation scheme in Table 3. The uniform boundary approach, dividing the entire video into segments based on the number of captions, results in moderate per-

♣ Co-first authors. Work done as interns at Shanghai AI Laboratory.
✉ Corresponding author.

Index	Prompt Description
1	Task: Extract ordered steps from video subtitles, focusing on sequential process and action-oriented instructions.
2	Task: Please extract concise and action-oriented steps from video subtitles, ensuring a clear sequential arrangement of events.
3	Task: Generate steps from subtitles in a single-sentence, emphasizing clear actions, excluding non-action-related events.
4	Task: Extract concise and action-oriented steps or instructions from the video subtitles, focusing solely on the sequential process. Each step should be presented as a single sentence with clear actions.
5	Task: Extract concise and action-oriented steps or instructions from the video subtitles, focusing solely on the sequential process. Each step should be presented as a single sentence with clear actions. Exclude any steps that are not directly related to the actions in the video.

Table 1. Examples of enhanced prompts for extracting step-by-step events from video subtitles.

Dataset	Data	M	C	S	Rec.	Pre.	F1
YouCook2	0%	7.87	46.02	6.87	29.66	42.04	34.78
	25%	8.64	52.47	7.89	**31.87**	43.25	36.70
	50%	9.14	55.42	7.37	29.02	44.87	35.25
	75%	9.39	**60.95**	**8.17**	31.27	44.61	**36.77**
	100%	**9.41**	59.35	7.97	30.80	**45.13**	36.61
ActivityNet	0%	8.31	30.11	5.63	50.82	55.73	53.16
	25%	8.85	30.11	5.63	51.18	56.29	55.60
	50%	8.72	30.00	5.70	53.29	57.93	55.51
	75%	8.84	**32.75**	**5.90**	**53.86**	56.67	55.23
	100%	**8.93**	31.89	5.85	53.14	**58.31**	**55.61**

Table 2. Impact of varying pretraining data on dense video captioning performance on YouCook2 and ActivityNet.

Boundary	M	C	S	Rec.	Pre.	F1
Uniform	2.95	15.16	4.29	19.4	19.88	19.64
Drop-DTW	3.21	17.37	3.86	14.09	17.26	15.51
Ours w/o STC	5.43	25.57	4.47	18.05	32.83	23.29
Ours w/ STC	**5.88**	**28.04**	**4.47**	**19.72**	**35.43**	**25.34**

Table 3. Comparative analysis of pseudo boundary generation schemes for dense video captioning on YouCook2.

formance. Surprisingly, the Drop-DTW method exhibits higher caption metric performance than uniform boundaries but notably underperforms in localization metrics. This discrepancy suggests that the noise present in the similarity matrix might hinder the accurate generation of boundaries using Drop-DTW. Conversely, our proposed scheme showcases superior performance in both localization and caption metrics, indicating the efficacy of our method in mitigating the impact of noise and improving overall performance in dense video captioning tasks.

Pretrain	FT Data	M	C	S	Rec.	Pre.	F1
✗	25%	5.06	22.39	4.45	23.94	36.41	28.89
✓	25%	7.81	46.69	7.17	28.93	41.08	33.95
✗	50%	6.45	34.30	5.44	25.18	39.75	30.83
✓	50%	8.60	55.73	7.84	30.14	42.73	35.35
✗	75%	7.26	40.49	6.31	27.81	40.60	33.01
✓	75%	9.11	59.09	7.86	29.97	44.54	35.83
✗	100%	7.87	46.02	6.87	29.66	42.04	34.78
✓	100%	**9.41**	**59.35**	**7.97**	**30.80**	**45.13**	**36.61**

Table 4. Impact of fine-tuning data and pretraining on few-shot dense video captioning performance in YouCook2. "FT data" represents the percentage of data used for fine-tuning.

Few-shot Model Performance In our investigation of few-shot dense video captioning, we analyze model performance with varying amounts of fine-tuning data on the YouCook2 dataset, as detailed in Table 4. Notably, as expected, models exhibit improved performance as more fine-tuning data is utilized, showcasing a positive correlation between data volume and model performance. Moreover, our comparison between models with and without pretraining reveals a substantial performance boost in pretraining-enabled models across all fine-tuning data proportions. Intriguingly, models with pretraining showcase significantly superior performance even with just half the fine-tuning data, surpassing the performance of models without pretraining. These results underscore the remarkable effectiveness of pretraining and our pseudo annotation generation method in enhancing model capabilities for few-shot dense video captioning tasks.

Effects of top-$\hat{k}$ in Pseudo Boundary Generation In our investigation of top-$\hat{k}$ values' impact on pseudo boundary generation and subsequent model performance on YouCook2 and ActivityNet, detailed in Table 5, contrasting trends emerge between the two datasets. Surprisingly, YouCook2 displays improved performance as the top-$\hat{k}$

Dataset	$\hat{k}$	M	C	S	Rec.	Pre.	F1
YouCook2	15	**5.91**	**30.89**	**5.00**	**20.56**	**35.92**	**26.15**
	20	5.79	27.04	4.72	18.99	35.89	24.84
	25	5.35	25.86	4.71	17.8	33.24	23.18
	30	4.62	21.88	4.48	17.26	30.05	21.93
ActivityNet	15	7.21	**18.40**	**4.76**	**43.96**	61.50	51.27
	20	7.30	17.52	4.69	43.46	63.68	51.66
	25	7.32	17.25	4.71	43.22	64.74	**51.84**
	30	**7.45**	16.50	4.71	42.29	**65.68**	51.45

Table 5. Impact of top-$\hat{k}$ values on model performance in YouCook2 and ActivityNet.

Dataset	K	M	C	S	Rec.	Pre.	F1
YouCook2	3	5.90	29.62	4.96	20.44	**36.38**	**26.17**
	4	5.89	29.28	4.85	19.85	35.67	25.51
	5	**5.91**	**30.89**	**5.00**	**20.56**	35.92	26.15
ActivityNet	3	7.33	15.86	4.67	42.21	**65.79**	51.43
	4	**7.45**	**16.50**	**4.71**	42.29	65.68	51.45
	5	7.35	16.27	4.68	**42.77**	65.54	**51.76**

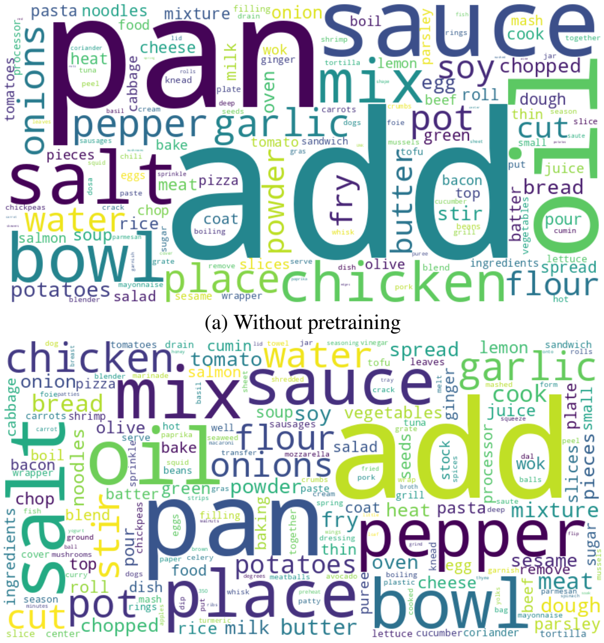
Table 6. Impact of selected proposals on boundary merging in YouCook2 and ActivityNet.

value decreases, while ActivityNet exhibits a contrasting trend, showing deteriorating results with smaller top-$\hat{k}$ values for partial metrics. Our pseudo boundary generation scheme operates such that a larger top-$\hat{k}$ value leads to an inclusion of more frames, potentially unrelated to the current event caption. We hypothesize that YouCook2, known for its diverse events, likely has a higher frequency of events, thereby resulting in shorter event durations compared to ActivityNet. This difference in event duration might explain the disparate impact of top-$\hat{k}$ values on model performance between the two datasets.

Effects of Refinement Hyperparameters. In Table 6, our exploration of the impact of the number of selected proposals for merging new boundaries (3, 4, and 5 proposals) reveals an intriguing trend across both YouCook2 and ActivityNet. As the number of proposals increases, we observe a trade-off in metrics. Fewer proposals tend to select boundaries closely aligned with the current caption, suggesting a focus on relevance, while a larger count encompasses a more diverse selection, albeit with some proposals less related to the ongoing caption. This dual impact leads to a compromise in model performance, highlighting the detrimental effects when the number of proposals deviates too far from an optimal range.

Dataset	Stages	M	C	S	Rec.	Pre.	F1
YouCook2	1	**5.95**	**32.08**	4.78	19.39	35.79	25.15
	2	5.90	29.62	4.96	**20.44**	**36.38**	**26.17**
	3	5.74	28.95	**4.98**	20.29	34.72	25.61
ActivityNet	1	7.34	15.93	4.65	**43.10**	65.5	**51.99**
	2	7.33	15.86	**4.67**	42.21	**65.79**	51.43
	3	**7.37**	**16.04**	4.65	42.40	65.39	51.44

Table 7. Impact of refinement stages on model performance in YouCook2 and ActivityNet.



(a) Without pretraining

(b) With pretraining

Figure 1. Word cloud of predicted captions from our DIBS w/ and w/o pretraining. A larger word indicates a higher frequency of its occurrence in the predicted captions.

In Table 7, we examine the influence of refinement stages (1, 2, and 3 stages) and observe distinct trends across both YouCook2 and ActivityNet. On YouCook2, we observe a trend of declining caption performance with increasing refinement stages, accompanied by a marginal decrease in localization metrics with larger stage counts. Similarly, ActivityNet exhibits varying metric shifts with increased stages, suggesting that an excessive number of refinement stages may not consistently enhance performance and could lead to fluctuations in metric outcomes. This nuanced relationship emphasizes that excessive refinement may not always yield improved results and could elongate the training process.

3.1. Visualization Results

Does pretraining lead to diverse captioning? In Figure 1, we compare word clouds of predicted captions from DIBS model with and without our proposed pretraining. It can be seen that predicted captions after pretraining (Figure 1b) show a richer vocabulary and broader concepts than those without pretraining (Figure 1a), highlighting our method's success in diversifying caption generation. The variation in word size and diversity after pretraining suggests significant improvements in the model's nuanced understanding, supporting our hypothesis on the benefits of pretraining for rich and diverse dense video captioning.

Qualitative Results Finally, to further prove the effectiveness of our method, we present some qualitative results of DIBS on YouCook2 and ActivityNet datasets in Figure 2 and Figure 3, respectively. We can see from these figures that our DIBS could predict accurate event boundaries along with rich captions.

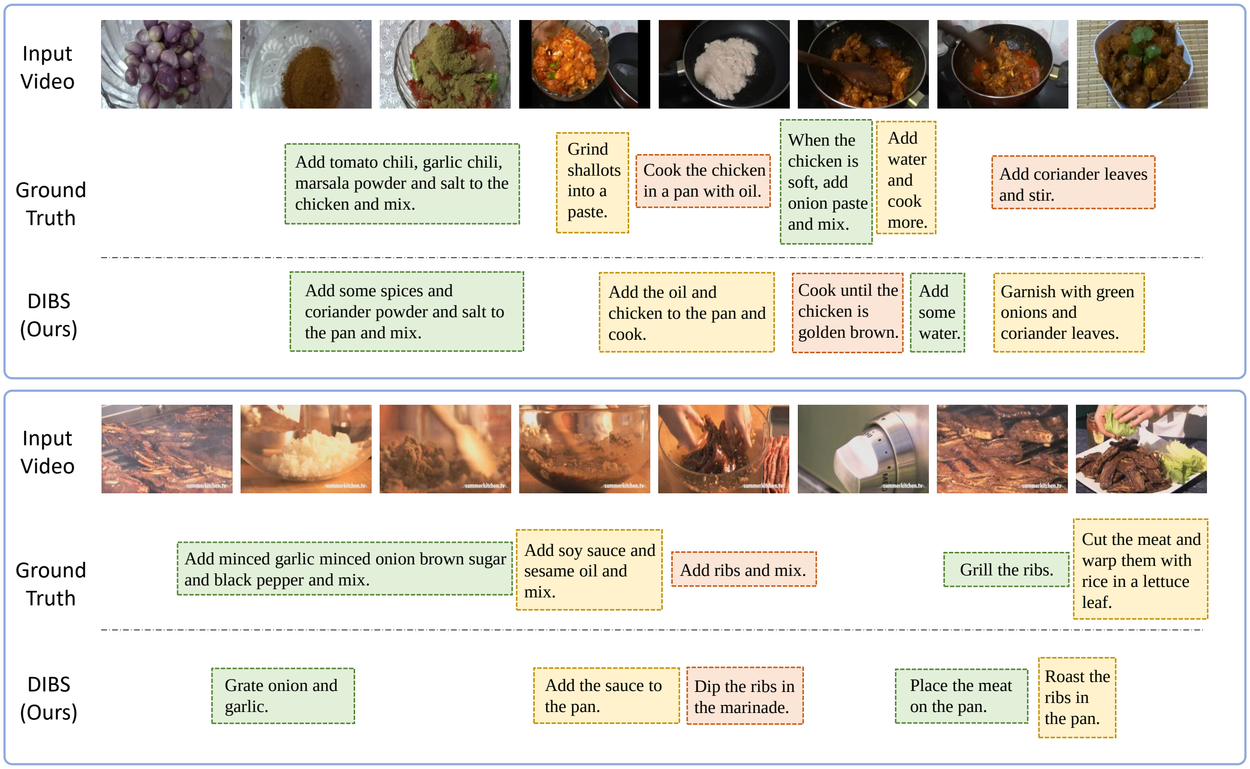
References

[1] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, 2017. 1

[2] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. 1

[3] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1

[4] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, pages 7590–7598, 2018. 1



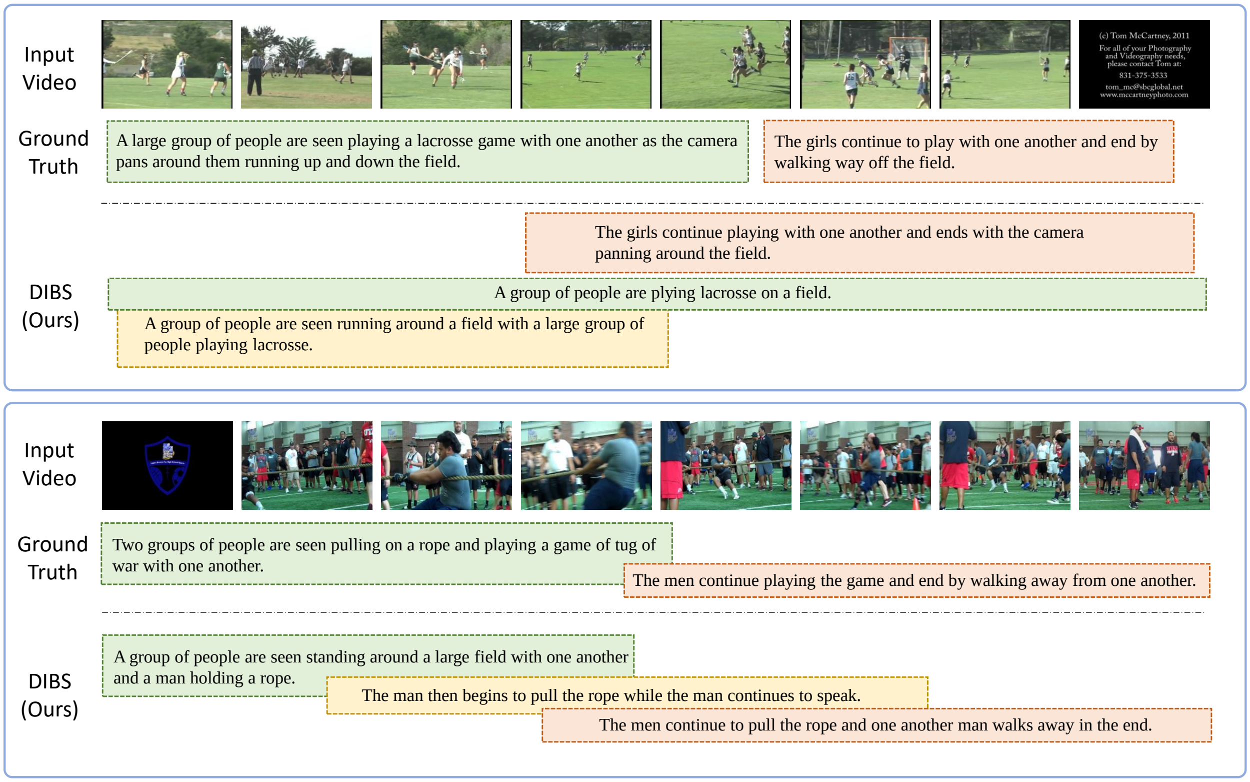
Ground Truth: Add tomato chili, garlic chili, marsala powder and salt to the chicken and mix. Grind shallots into a paste. Cook the chicken in a pan with oil. When the chicken is soft, add onion paste and mix. Add water and cook more. Add coriander leaves and stir.

DIBS (Ours): Add some spices and coriander powder and salt to the pan and mix. Add the oil and chicken to the pan and cook. Cook until the chicken is golden brown. Add some water. Garnish with green onions and coriander leaves.

Ground Truth: Add minced garlic minced onion brown sugar and black pepper and mix. Add soy sauce and sesame oil and mix. Add ribs and mix. Grill the ribs. Cut the meat and warp them with rice in a lettuce leaf.

DIBS (Ours): Grate onion and garlic. Add the sauce to the pan. Dip the ribs in the marinade. Place the meat on the pan. Roast the ribs in the pan.

Figure 2. Visualization of qualitative results of our DIBS model. Video examples are selected from the validation set of YouCook2.



Ground Truth: A large group of people are seen playing a lacrosse game with one another as the camera pans around them running up and down the field. The girls continue to play with one another and end by walking way off the field.

DIBS (Ours): The girls continue playing with one another and ends with the camera panning around the field. A group of people are plying lacrosse on a field. A group of people are seen running around a field with a large group of people playing lacrosse.

Ground Truth: Two groups of people are seen pulling on a rope and playing a game of tug of war with one another. The men continue playing the game and end by walking away from one another.

DIBS (Ours): A group of people are seen standing around a large field with one another and a man holding a rope. The man then begins to pull the rope while the man continues to speak. The men continue to pull the rope and one another man walks away in the end.

Figure 3. Visualization of qualitative results of our DIBS model. Video examples are selected from the validation set of ActivityNet.