APPENDIX FOR ACTIONS INSPIRE EVERY MOMENT: ONLINE ACTION-AUGMENTED DENSE VIDEO CAPTIONING

Anonymous authors
Paper under double-blind review

A APPENDIX

A.1 ADDITIONAL IMPLEMENTATION DETAILS

We utilize the CLIP model pretrained on the LAION-2B dataset. Specifically, we use its ViT-Large model (303M parameters) and its 12-layer Transformer text model (128M parameters). The CLIP-initialized ViT is kept frozen throughout all stages of training described below.

We further pretrain the CLIP text model on the image captioning task using the same LAION-2B dataset, with batch size 1024 for 0.2 epochs. We use the Adam optimizer with momentum 0.9, an initial learning rate (LR) of 5e-5, 5000 warmup steps, linear LR decay, weight decay 1e-2.

For dense video captioning, the token reduction Transformer and autoregressive Transformer modules are added. Each module consists of 8 layers with a model dimension of 512, and 32 M parameters. In total, the entire model contains approximately 500M parameters.

As described in section 3.3, we apply image-based simulated video pretraining on the entire model, including the newly added modules. To simulate video sequences, we sample 3 to 5 images from the LAION-2B dataset, repeating each image multiple times to form a 16-frame sequence. To create smoother transitions, we blend pixels at the boundaries by applying a weighted sum of two images, using a randomly selected blending ratio $\alpha \in [0.1, 0.9]$, *e.g.*, blending pixels of images A and B as $\alpha A + (1 - \alpha)B$. We apply random augmentations to each frame to avoid overly monotonous sequences. This pretraining follows the same segment-by-segment autoregressive framework for online dense video captioning. We use a batch size of 32 and train the model for 100000 steps. The optimizer is Adam with momentum 0.9, an initial LR of 1e-4, 5000 warmup steps, cosine LR decay, and a weight decay of 1e-5.

When finetuning on dense video captioning, the model is trained for 20000 steps with a batch size 16. We again use the Adam optimizer with momentum 0.9, an initial LR of 1e-4, 5000 warmup steps, cosine LR decay and a weight decay of 1e-5. For time tokenization, we use relative time tokens following Vid2Seq (Yang et al., 2023). We quantize a video of duration T frames into $B = 32$ equally spaced time bins.

For inference, we follow the standard protocol to use beam search, with a beam size of 6 and temperature 1, followed by temporal NMS with a threshold of 0.7 to remove overlapping intervals.

A.2 COMPUTATIONAL COST OF OUR MODEL

Our model uses 410 GFLOPs per segment, totaling 6560 GFLOPs for 16 segments. The retrieval operates on the precomputed text embeddings.

A.3 PROMPTING FOR ACTION TEXT CORPUS CONSTRUCTION

We construct a corpus of action phrases to serve as contextual priors during the dense video captioning process. These phrases are designed to capture key actions or events in each video segment and identify relevant objects. To build this corpus, we draw from two main sources: 1) Captions from the training splits of dense video captioning datasets: ViTT, YouCook2, and ActivityNet, where we use only the text captions excluding the video frames. 2) A broader, less domain-specific corpus

from the HowTo100M dataset (Miech et al., 2019), again using only the subtitle text without the video content.

Unlike previous methods that use raw video captions, which tend to be lengthy and unfocused (Xu et al., 2024), we summarize these captions into concise action phrases using the publicly available language model, Gemma (Team et al. (2024), huggingface.co/google/gemma-2-27b).

To improve the extraction process, we refine the prompt to ensure concise and consistently formatted action phrases. Specifically, we emphasize singular nouns, avoid numerical terms, and enforce a strict format of <action verb(ing)> <target object (if any)>. For example, our prompt is: *Your goal is to summarize the input sentence using as few words as possible. Focus on the words describing actions or events. Use singular nouns, avoid articles and numeric terms. Respond in the format of <action verb (ing)> <target object (if any)>. Input: {raw caption}. Answer:*

For HowTo100M subtitles, which are often longer, we adjust the prompt to focus on extracting a single main action or event: *The input is video subtitle text. Choose the main action or event in the video and summarize it using as few words as possible. Focus on the words describing actions or events. Use singular nouns, avoid articles and numeric terms. Respond in the format of <action verb(ing)> <target object (if any)>. Input: {video subtitles}. Answer:*

These prompts effectively generate concise action phrases. After processing all text in each source, we deduplicate phrases by merging those with the same set of words, regardless of word order. After filtering out some least frequent phrases, we obtain 30,000 action phrases for each corpus.

A.4 Additional Ablations

We conduct additional ablation studies using the same setup described in section 4.2, where we use 16 frames per video, with 1 frame per segment at a resolution of 256×256 pixels, and report the results on the ViTT dataset. In this ablation, the mixed training (Table 6) and image-based simulated video pretraining (table 8) are *not* used, unless otherwise noted.

Effect of text decoder pretraining. In Table A1, we show the effect of pretraining the text decoder for image captioning using the LAION-2B dataset (section 3.1). While image captioning pretraining improves performance, our model still performs reasonably well without it. Notably, most recent dense video captioning methods (Yang et al., 2023; Wang et al., 2024; Ren et al., 2024; Wu et al., 2024; Zhou et al., 2024) employ language pretraining for their text decoders, and we follow this approach to enhance the captioning performance.

Ablation on the size of action text corpus. Table A2 presents the effect of varying the size of the action text corpus. We randomly sample subsets of the corpus at 1%, 10%, 50%, and 100%. Increasing the corpus size improves performance, with the most notable gains observed up to 50%. This shows that our constructed corpus is effective in covering a broad range of action phrases for retrieval augmentation.

Ablation on the number of segments. Table A3 studies the effect of the number of segments. Overall, more frames improves performance. Across different segment configurations, our model performs robustly overall, with 16 segments yielding the best results.

Ablation on ViTT, YouCook2, ActivityNet datasets. In Table A4, we evaluate the effects of our key method components. The baseline refers to the online model described in section 4.1. We observe both our main contributions – online action-augmentation (section 3.2) and image-based simulated video pretraining (section 3.3) – make complimentary improvements to performance across all three benchmarks: ViTT, YouCook2, and ActivityNet.

A.5 Comparison of approaches in existing methods

Table A5 compares various strategies used in existing methods, focusing on key aspects such as support for online video captioning, reliance on video-text pretraining, and the backbone models employed.

text decoder pretraining	S	C	M
✓	9.9	37.2	9.6
x	8.3	30.4	8.0

Table A1: **Ablation on text model pretraining** on the image captioning task. Results on ViTT.

size of action text corpus	S	C	M
1%	8.7	33.0	8.4
10%	9.3	35.2	9.1
50%	9.7	36.7	9.5
100%	9.9	37.2	9.6

Table A2: **Effect of size of action text corpus.** Results on ViTT.

# segments	# frames	S	C	M
8	8	9.3	36.2	8.9
8	16	9.6	37.0	9.1
16	16	9.9	37.2	9.6
16	32	10.0	37.7	9.8
32	32	9.7	36.8	9.0

Table A3: **Number of segments** which controls the number of decoding outputs. Results on ViTT.

	ViTT				YouCook2				ActivityNet			
method	S	C	M	F1	S	C	M	F1	S	C	M	F1
baseline online model	7.6	27.7	7.2	34.0	5.3	27.7	6.8	22.4	5.4	31.6	9.8	45.8
online action-augmentation	9.8	37.4	9.4	35.5	6.9	39.4	8.0	25.5	6.7	34.6	11.2	46.2
image-based simulated video pretraining	8.9	34.1	8.5	37.8	6.1	35.0	7.2	27.8	6.2	33.7	10.4	47.6
both combined	10.8	39.1	10.3	39.2	8.0	45.6	9.3	30.7	7.5	36.4	12.1	49.9

Table A4: **Ablation of our method on ViTT, YouCook2, ActivityNet datasets.** This ablation uses S=16 segments per video, L=1 frame per segment at a resolution of 256×256 pixels.

method	online	video-text pretraining	backbone
E2ESG (Zhu et al., 2022)	N	∅	C3D
PDVC (Wang et al., 2021)	N	∅	TSN
OmniViD (Wang et al., 2024)	N	Kinetics	VideoSwin + Bart
TimeChat (Ren et al., 2024)	N	YT-Temporal, ViTT, ActivityNet, etc.	Eva-CLIP-G + Llama-7B
Vid2Seq † (Yang et al., 2023)	N	YT-Temporal-1B	CLIP-L + Bert-B
DoYou (Kim et al., 2024)	N	∅	CLIP-L
DIBS (Wu et al., 2024)	N	Howto100M	CLIP-L
Streaming (Zhou et al., 2024)	**Y**	YT-Temporal-1B	CLIP-L + Bert-B
AIEM (ours)	**Y**	∅	CLIP-L

Table A5: **Comparison to the state-of-the-art on dense video captioning.**

REFERENCES

Minkuk Kim, Hyeon Bae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. Do you remember? dense video captioning with cross-modal memory retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13894–13904, 2024.

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.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14313–14323, 2024.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Junke Wang, Dongdong Chen, Chong Luo, Bo He, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. Omnivid: A generative framework for universal video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18209–18220, 2024.

Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *ICCV*, 2021.

Hao Wu, Huabin Liu, Yu Qiao, and Xiao Sun. Dibs: Enhancing dense video captioning with unlabeled videos via pseudo boundary enrichment and online refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18699–18708, 2024.

Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13525–13536, 2024.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. *CVPR*, 2023.

Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18243–18252, 2024.

Wanrong Zhu, Bo Pang, Ashish V. Thapliyal, William Yang Wang, and Radu Soricut. End-to-end dense video captioning as sequence generation. In *COLING*, 2022.