Multi-Modal Few-Shot Temporal Action Segmentation

Supplementary Material

In our paper, we introduce the new problem of Multi-Modal Few-Shot Temporal Action Segmentation (MMF-TAS). MMF-TAS imposes unique challenges–generalizing to novel tasks with a few support videos and exploiting information of both visual and textual modalities. We propose the first MMF-TAS framework with our Prototype Graph Network (PGNet). In this supplementary material, we provide more implementation details and ablation studies.

1. Implementation Details

PGNet. As mentioned in the paper, we use the video and text encoders of ProceduralVRL [12]. As their training set [4] and CrossTask and COIN are collected from Youtube, we have checked and ensured there is no overlapping between their training set and our test videos. To compute textual features of action classes from the text encoder, we use the 28 prompts defined in [12], such as *"A video of a person doing <action name>", "A example of a person doing <action name>"*, etc. For each class, we use the average of the features of all prompts. Meanwhile, to adapt the video encoder to our long video datasets, we follow the side tuning method [8] that uses a temporal convolution based on its last-layer's feature to improve its temporal modeling ability.

Recall that we define our fusion function ψ as a weighted combination of attentions in Eq(9) of the paper, $\psi(\hat{\mathbf{\Delta}}_l, \mathbf{\Delta}_{l-1}) = \hat{\mathbf{\Delta}}_l + \tau_l \mathbf{\Delta}_{l-1}$. Functions like convolutions or transformers are not applicable as ψ, as they require inputs with at least one fixed dimension. As $\hat{\mathbf{\Delta}}_l$ and $\mathbf{\Delta}_{l-1}$ are the attentions between the query video and action prototypes, we have $\hat{\mathbf{\Delta}}_l, \mathbf{\Delta}_{l-1} \in \mathbb{R}^{L \times A}$. L is the query video length and A is the number of actions in the tasks, both of which change as query/support videos change. Our experimental results have shown that weighted combination is a simple yet effective solution to fuse predictions while also allowing an influence factor τ_l to prevent over-reliance on one modality.

VLM baselines. We use VLMs as our zero-shot baselines and provide here their setup details. For contrastive VLMs, we obtain textual features of action names with their text encoders. We split videos into short clips and compute per-clip features with their video encoders. We then classify the actions in clips by the similarity between the visual and textual features. We finetune their last layers on our datasets as full-parameter finetune causes overfitting and lower zero-shot results.

For generative VLMs, they do not have a text or video encoder but use LLM to generate text-form outputs. We first test passing a video clip and a list of action names as input and instructing them to choose actions from the list. However, it leads to poor performance as models do not always follow the instruction. Thus, we use them to get caption for each clip and compute the similarity between the caption and action names via a sentence embedding method [6]. As it is too costly to caption each clip, we first leverage the unsupervised method in [7] to partition a video into potential action segments, then generate a caption for each clip. Lastly, generative VLMs also fail to detect background frames, which are frames not relevant to any action. We apply thresholding on the similarity score between clip captions and action names, to choose frames with low scores as background frames. As this threshold is tuned on the videos of base tasks, it leads to higher performance on base tasks than novel tasks.

2. Experimental Results

In Table 1, we show the comprehensive results of our PGNet with the state-of-the-art methods on CrossTask and COIN datasets. For each zero-shot and few-shot setting, we include the results of both base and novel tasks, in the format of **novel task**|**base task**. We especially include a new setting of Few-Shot *All*way-3shot, where models adapt to all 6 novel tasks in CrossTask and all 60 novel tasks in COIN. We highlight the performance improvement of our PGNet on novel tasks in blue, which shows significant improvement over the best baseline in all settings.

3. Few-Shot Evaluation with Mixed Tasks

In Table 1 of the paper, we test models with query and support videos sampled from K tasks. The tasks are either base or novel tasks. We now evaluate with a more challenging setting where K tasks are a mixture of base and novel tasks. In this case, models often will match query videos to actions from base tasks and under-predict those from novel tasks [10].

Recall that PGNet uses visual predictions for novel tasks and textual predictions for base tasks. In the previous setting, we choose the predictions to use by detecting if all support videos match with the base tasks in training data or not. In the current setting, support videos are sampled from both base and novel tasks. Thus, we leverage the task prediction, $argmax(\mathbf{P}^v_{task} + \mathbf{P}^t_{task})$, to match query videos to one of the K tasks. Then we can detect if the task is a base or novel task by comparing its support videos with training videos, and use the corresponding predictions.

As shown in Table 3, we observe negligible performance

	CrossTask					COIN				
	Edit	F1@10	F1@25	F1@50	Acc	Edit	F1@10	F1@25	F1@50	Acc
Zero-Shot 3way-3shot										
ProceduralVRL[12]	13.7|14.6	13.0|13.5	9.5|10.7	5.8|6.5	30.8|42.6	12.6|19.2	14.0|15.5	10.8|11.5	5.3|6.2	34.7|39.9
LanguageBind[13]	11.4|15.4	10.4|16.3	7.8|13.3	4.3|7.9	20.1|50.0	12.1|15.1	11.3|13.4	8.9|10.8	4.9|5.6	35.6|42.1
Chat-UniVi[1]	9.5|9.9	7.9|9.0	6.2|7.2	3.8|4.1	9.2|10.3	9.4|10.2	6.9|8.9	5.4|6.1	2.3|3.8	6.3|9.3
PGNet *(zero-shot)*	**36.1|54.2**	**32.2|57.9**	**26.4|51.2**	**15.4|33.9**	**36.4|62.7**	**31.8|59.4**	**24.0|57.3**	**17.4|49.3**	**8.6|31.3**	**36.6|63.8**
	+22.4	*+19.2*	*+16.9*	*+9.6*	*+5.6*	*+19.2*	*+10.0*	*+6.6*	*+3.3*	*+1.0*
Few-Shot 3way-3shot										
Linear Probe	15.4|20.8	14.8|19.1	11.0|15.2	6.0|9.1	31.4|43.1	12.6|19.2	18.5|23.5	14.2|17.1	7.5|9.4	34.6|39.8
MUPPET[5]	18.1|29.4	19.5|28.0	15.8|24.1	9.5|14.6	13.1|24.0	13.1|18.8	19.9|31.0	15.4|24.4	8.1|13.4	15.9|22.3
PGNet *(no-label)*	35.8|54.2	33.4|58.2	27.5|51.4	16.9|34.3	35.9|62.5	37.2|59.0	27.8|57.0	19.9|49.1	10.2|31.1	29.6|62.5
PGNet *(weak-label)*	36.5|54.8	35.4|58.4	29.1|51.5	17.3|34.1	37.6|63.2	46.1|59.1	35.3|57.3	26.1|49.7	13.8|31.4	42.4|63.2
PGNet *(full-label)*	**37.7|54.8**	**36.6|58.4**	**30.1|51.6**	**18.3|34.2**	**39.3|63.2**	**46.6|59.3**	**36.7|57.3**	**27.6|49.3**	**15.0|31.3**	**43.8|63.8**
	+19.6	*+17.1*	*+14.3*	*+8.8*	*+7.9*	*+33.5*	*+16.8*	*+12.2*	*+6.9*	*+9.2*
Few-Shot 5way-3shot										
Linear Probe	13.4|12.8	12.7|20.8	9.8|15.4	5.4|7.9	27.2|41.9	13.0|16.5	15.6|21.3	11.5|15.5	6.3|8.2	33.1|39.1
MUPPET [5]	15.8|28.0	17.3|27.6	13.7|23.2	8.6|13.9	12.5|27.2	11.5|17.0	19.2|29.0	14.5|22.8	7.6|12.1	14.2|20.7
PGNet *(no-label)*	33.2|51.6	31.9|55.8	26.6|49.7	16.4|31.8	33.4|61.4	34.4|59.4	27.4|58.3	19.9|49.5	9.8|32.9	29.6|63.2
PGNet *(weak-label)*	34.2|51.9	33.3|56.1	28.2|50.0	17.4|32.1	35.8|62.4	44.6|59.7	34.9|58.5	26.1|49.1	13.1|32.3	41.9|62.8
PGNet *(full-label)*	**34.7|51.7**	**33.7|56.1**	**28.2|50.0**	**17.5|32.1**	**35.7|62.4**	**45.4|59.6**	**36.2|58.2**	**27.5|49.2**	**14.3|32.8**	**43.3|63.5**
	+18.9	*+16.4*	*+14.5*	*+8.9*	*+8.5*	*+32.4*	*+17.0*	*+13.0*	*+6.7*	*+10.2*
Few-Shot *All*way-3shot										
Linear Probe	10.2|12.3	13.1|14.6	8.6|10.8	4.4|5.5	26.9|38.6	8.2|10.6	9.4|12.0	7.0|8.8	3.8|4.7	25.2|28.3
MUPPET [5]	14.2|24.2	14.5|25.2	11.2|21.2	7.7|12.8	12.1|23.3	10.7|15.4	11.7|14.3	8.6|9.8	3.9|5.0	15.7|21.5
PGNet *(no-label)*	31.4|52.1	29.8|56.7	24.4|50.5	14.6|33.6	31.5|61.5	20.2|52.9	15.4|51.2	11.0|43.8	5.7|28.6	14.8|53.9
PGNet *(weak-label)*	32.1|53.0	31.5|57.0	25.3|50.8	15.7|33.7	31.6|62.1	27.1|52.4	20.7|50.9	15.1|43.9	7.6|28.3	20.7|53.0
PGNet *(full-label)*	**34.6|53.7**	**33.5|57.8**	**27.8|51.5**	**17.0|34.3**	**33.8|62.8**	**40.8|59.4**	**31.9|57.9**	**22.4|49.3**	**11.5|31.8**	**38.1|63.1**
	+20.4	*+19.0*	*+16.6*	*+9.3*	*+6.9*	*+30.1*	*+20.2*	*+13.8*	*+7.6*	*+12.9*

Table 1. **Performance on CrossTask and COIN datasets**. We test with four settings and report results on both base and novel tasks in the format of **novel task**|**base task**. In each setting, we show in blue the PGNet's improvement on novel tasks over the best baseline.

loss, due to the fact that our Action Relation Graph includes task nodes to identify and distinguish each task. It enables accurate prediction of the task labels for query videos and achieves a prediction accuracy of 97%. Thus, we can infer action labels with action prototypes of the correct task.

4. Choice of VLM Backbone

In this section, we study the effect of using different VLMs as our encoder. As discussed in the paper, there are two mainstreams of VLMs – contrastive and generative VLMs, each having unique advantages and disadvantages. Contrastive VLMs learn shared semantic spaces to compare video and text features, thus are directly applicable to MMF-TAS. On the other hand, generative VLMs employ LLMs to produce video captions that can contain rich information about the input while also allowing user-given prompts to guide the caption. Yet they lack explicit semantic spaces to compare videos and texts.

Implementation. To compute video embedding via generative VLM, we apply unsupervised partition method [7] to split a video into clips. For each clip, we pass it into VLM with the prompt of *"This is a clip from a procedural video. Give a short description of the action of the person in this clip with this template: The person is doing ..."*, which we found leads to concise and accurate clip captions. Next, we follow recent work [11] of LLM sentence embedding and use the penultimate layer's features of generated captions as features of the video clips. These features have been shown to contain rich information about the inputs. Similarly, we compute textual features of actions with the text prompt of *"Performing [task name] requires action [action name]. Give a short description of the action."*

Experimental Results. Table 2 shows the effectiveness of different VLMs on CrossTask with few-shot 3way-3Shot setting. It can be observed that, the contrastive VLMs achieve better performance than the generative VLM. It shows their encoders provide fine-grained and aligned feature encodings, easily support varied downstream tasks. ProceduralVRL has better performance than LanguageBind likely because its pretraining dataset is Howto100M [4], which contains procedural videos similar to CrossTask and COIN datasets. For generative VLM, interestingly, we found the main reason for its low performance is model hallucination. It often gives wrong, confident action descriptions even though the video clips do not contain any meaningful or observable action. This underscores a chal-

	Edit	F1@10	F1@25	F1@50	Acc
ProceduralVRL[12]	37.7|54.8	36.6|58.4	30.1|51.6	18.3|34.2	39.3|63.2
LanguageBind[13]	30.4|54.5	27.9|57.9	23.8|52.2	15.8|36.6	31.9|64.1
Chat-UniVi[1]	34.8|48.9	28.5|49.2	21.9|40.3	12.0|24.0	29.9|56.0

Table 2. **Effect of Different Video-Language Model as Video/Text Encoders** with few-shot 3way-3shot setting on CrossTask.

		Edit	F1@10	F1@25	F1@50	Acc
Mixed	3way-3shot	45.3|61.2	36.2|58.0	27.3|49.7	14.6|32.0	43.7|62.2
Separate		**46.6|59.3**	**36.7|57.3**	**27.6|49.3**	**15.0|31.3**	**43.8|63.8**
Mixed	5way-3shot	45.3|58.8	35.7|57.2	26.7|49.5	13.9|32.1	40.0|64.8
Separate		**45.4|59.6**	**36.2|58.2**	**27.5|49.2**	**14.3|32.8**	**43.3|63.5**

Table 3. **Evaluation with Mixtures of Base and Novel Tasks** for few-shot settings on COIN.

lenge in caption-based video understanding: while models can produce rich description about videos, they may not accurately capture the information needed by the users. Hence, classification-based video understanding allows explicit defining the actions of interest with action labels.

5. Generalization with More Support Videos.

In Table 4, we compare the effect of using different numbers of support videos for few-shot learning, ranging from one video per task to five videos. As expected, including more videos provides richer visual demonstration, thus improving model performance. However, the long durations of procedural videos create difficulty in both collecting high-quality videos and fully annotating them. Thus, we assume 3 support videos per task (Kway-3shot) in most experiments to ensure low data curation cost.

Additionally, we train the latest fully-supervised TAS method [3] on CrossTask and COIN. This method is trained on both base and novel tasks and evaluated on all tasks, yielding {F1@10, F1@25, F1@50} scores of {56.6, 50.8, and 35.1} on CrossTask and {51.7, 45.3, and 30.8} on COIN, respectively. Although its performance is not directly comparable to that of PGNet due to different problem settings, it highlights the challenging nature of CrossTask and COIN [2, 9, 14]. They comprise YouTube videos featuring diverse environments, appearance variations, viewpoint changes, and a high proportion of frames that are irrelevant to the task, among other factors. Despite these challenges, PGNet achieves promising results using only three support videos per task, demonstrating its ability to generalize to novel tasks at a low annotation cost.

	Edit	F1@10	F1@25	F1@50	Acc
3way-1shot	34.4|54.6	31.7|57.6	26.2|51.0	16.3|34.2	35.1|62.8
3way-3shot	37.7|54.8	36.6|58.4	30.1|51.6	18.3|34.2	39.3|63.2
3way-5shot	40.0|53.7	37.6|57.9	31.5|51.9	19.8|34.5	40.1|62.3

Table 4. **Effect of Varying the Number of Support Videos** for few-shot settings on CrossTask.

References

[1] Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*, 2023. 2, 3

[2] Z. Lu and E. Elhamifar. Weakly-supervised action segmentation and alignment via transcript-aware union-of-subspaces learning. *International Conference on Computer Vision*, 2021. 3

[3] Z. Lu and E. Elhamifar. Fact: Frame-action cross-attention temporal modeling for efficient action segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 3

[4] A. Miech, D. Zhukov, J. B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. *International Conference on Computer Vision*, 2019. 1, 2

[5] Sauradip Nag, Mengmeng Xu, Xiatian Zhu, Juan-Manuel Perez-Rua, Bernard Ghanem, Yi-Zhe Song, and Tao Xiang. Multi-modal few-shot temporal action detection via vision-language meta-adaptation. *arXiv preprint arXiv:2211.14905*, 2022. 2

[6] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019. 1

[7] Saquib Sarfraz, Naila Murray, Vivek Sharma, Ali Diba, Luc Van Gool, and Rainer Stiefelhagen. Temporally-weighted hierarchical clustering for unsupervised action segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2

[8] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Lst: Ladder side-tuning for parameter and memory efficient transfer learning, 2022. 1

[9] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[10] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1

[11] Bowen Zhang, Kehua Chang, and Chunping Li. Simple techniques for enhancing sentence embeddings in generative lan-

guage models. In *International Conference on Intelligent Computing*, pages 52–64. Springer, 2024. 2

[12] Yiwu Zhong, Licheng Yu, Yang Bai, Shangwen Li, Xueting Yan, and Yin Li. Learning procedure-aware video representation from instructional videos and their narrations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14825–14835, 2023. 1, 2, 3

[13] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Wang HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment, 2023. 2, 3

[14] D. Zhukov, J. B. Alayrac, R. G. Cinbis, D. Fouhey, I. Laptev, and J. Sivic. Cross-task weakly supervised learning from instructional videos. *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3