Decompose and Transfer: CoT-Prompting Enhanced Alignment for Open-Vocabulary Temporal Action Detection

Supplementary Material

In the supplementary material, we provide additional experiments to further substantiate the effectiveness of our proposed method. We include more implementation details, covering the extraction processes of visual and textual features, model architecture, and training configurations. For our core component, CoT-Prompting Semantic Decomposition (CSD), we conduct comprehensive ablations examining the influence of different prompt templates, varying phase numbers, and alternative commercial LLMs for label decomposition, as well as subjective evaluations of the generated descriptions using both GPT-based and human raters. Furthermore, we offer more empirical results demonstrating the plug-and-play flexibility of PDA, analyze the impact of diverse visual and textual backbones, compare against prior LLM-based label expansion strategies, report semantic similarity evaluations on the ActivityNet v1.3 dataset, and provide additional qualitative visualizations.

1. More Implementation Details

Following prior works [8, 12, 14], we employ a two-stream I3D model and the CLIP [13] model for feature extraction.
For visual features, we concatenate the RGB and optical flow features extracted from the two-stream I3D. On THUMOS14, video features are extracted from 16-frame segments using a sliding window with a stride of 4. On ActivityNet v1.3, features are extracted with a stride of 16 and subsequently downsampled to 128 dimensions.
For textual features, we use the frozen pre-trained CLIP text encoder (ViT-B/16 and ViT-L/14 variants).
For the model architecture, the temporal transformer consists of 6 layers, each comprising a multi-head self-attention (MHSA) followed by a feed-forward network (FFN), with a hidden dimension of 512. The weighting network W_p is implemented as a lightweight Transformer. We first replicate the input 512-dimensional visual feature into four virtual tokens, each representing a phase, and add four learnable phase embeddings to introduce phase-specific priors. The resulting tokens are processed by a 4-head MHSA layer. A linear projection with a hidden dimension of 1024 maps each token to a scalar, and a softmax operation produces the final 4-dimensional phase-wise weight vector. The linear projection layer L is implemented as a three-layer MLP with a hidden dimension of 1024.
During training, we adopt the Adam optimizer with a linear warm-up for the first 5 epochs. The initial learning rate is set to 0.0001. A MultiStepLR scheduler is applied for THUMOS14, while cosine annealing [11] is used for ActivityNet v1.3. The batch size is set to 16 for ActivityNet v1.3 and 2 for THUMOS14. The code has been submitted along with this pdf.

2. More Analysis of CoT-Prompting Semantic Decomposition

2.1. Analysis of the Prompt Template Design

In the main submission, we adopt a single prompt template to guide the chain-of-thought reasoning of GPT-4o for generating both phase-specific and global action descriptions. Here, we examine the robustness of our approach to prompt variations by introducing two additional templates, as summarized in Table 1. Although these prompts differ in surface phrasing, they share the same objective—eliciting coherent phase-aware descriptions or holistic motion summaries for each action. We evaluate the performance of our method using descriptions generated from each prompt variant. As shown in Table 2, all prompt versions yield comparable results, with only minor variations across evaluation metrics. This observation demonstrates two key aspects of robustness: First, GPT-4o shows strong insensitivity to prompt wording; despite differences in linguistic form, it consistently produces coherent and phase-aligned descriptions that capture the essential characteristics of each action. Second, our model is also robust to variations in the input descriptions themselves. Its performance is largely governed by the CoT-Prompting Semantic Decomposition (CSD) and Adaptive Phase-wise Alignment (APA) modules, which focus on learning transferable temporal patterns through phase-wise alignment rather than relying on subtle textual differences among prompt-generated descriptions.

2.2. Analysis of Different Phase Number on ActivityNet v1.3

In the main submission, we analyze the impact of different phase number on THUMOS14 dataset. Here, we provide more results on ActivityNet v1.3 dataset. As shown in Table 3, performance consistently improves as the number of phases increases, confirming that finer-grained decomposition provides richer semantic cues and benefits unseen action detection. However, the gains become marginal once the phase count exceeds four, reflecting the saturation of informative semantics and the emergence of redundant or noisy descriptions. Moreover, using more phases increases computational cost due to additional alignment operations. Balancing accuracy and efficiency, we therefore adopt a

Table 1. The variants of prompt templates used to guide GPT-4o in generating phase-specific and global action descriptions. The placeholder $\{text\}$ indicates the content to be filled in by GPT-4o.

Description	Type	Prompt
Phase-specific	**(a)**	***Question*: Given an action of $\langle Action \rangle$, considering how the activity typically begins, evolves, and concludes. After your reasoning, provide a concise phase-wise summary.** *Answer*: In the start phase, the person would $\{text\}$. In the middle phase, the person would $\{text\}$. In the end phase, the person would $\{text\}$.
	(b)	***Question*: For the given action of $\langle Action \rangle$, enumerate it into chronological sub-events and condense these events into three coherent phase descriptions.** *Answer*: In the start phase, the person would $\{text\}$. In the middle phase, the person would $\{text\}$. In the end phase, the person would $\{text\}$.
	(c)	***Question*: Decompose the action of $\langle Action \rangle$ into coherent three phases based on the natural temporal progression of the action. Please provide the output step by step.** *Answer*: In the start phase, the person would $\{text\}$. In the middle phase, the person would $\{text\}$. In the end phase, the person would $\{text\}$.
Global	**(a)**	***Question*: Describe the motion of a person does $\langle Action \rangle$.** *Answer*: The person would $\{text\}$.
	(b)	***Question*: Describe the motion of a person carries out $\langle Action \rangle$.** *Answer*: The person would $\{text\}$.
	(c)	***Question*: Describe how a person does $\langle Action \rangle$.** *Answer*: The person would $\{text\}$.

Table 2. Analysis of the Prompt Template Design. The prompt template adopted in this paper is highlighted with a blue background.

Data Split	Prompt Type	THUMOS14				ActivityNet v1.3			
		0.3	0.5	0.7	Avg	0.5	0.75	0.95	Avg
50% Seen 50% Unseen	**(a)**	64.9	49.4	24.0	46.4	52.9	35.1	7.7	34.3
	(b)	64.5	49.0	23.7	46.2	52.6	34.8	7.3	34.0
	(c)	65.4	49.7	24.3	46.9	53.1	35.3	7.7	34.6
75% Seen 25% Unseen	**(a)**	69.6	53.9	27.7	51.5	55.4	37.0	8.5	36.8
	(b)	69.3	53.5	27.4	51.3	55.1	36.9	8.3	36.6
	(c)	70.5	54.6	28.3	52.1	56.2	37.8	8.6	37.4

Table 3. Analysis of different action phase number on ActivityNet v1.3 under the 50% seen / 50% unseen split. The phase number adopted in this paper is highlighted with a blue background.

Phase Number	mAP@tIOU (%)				Time (min)
	0.3	0.5	0.7	Avg	
One *(Glob)*	51.0	32.8	5.9	32.3	14.7
Two *(Start, End)*	51.9	33.7	6.6	33.1	15.1
Three *(Start, Mid, End)*	52.6	34.5	7.1	33.9	15.8
Four *(Start, Mid, End, Glob)*	53.1	35.3	7.7	34.6	16.4
Five *(Start, Mid1, Mid2, End, Glob)*	53.5	35.5	8.0	34.8	17.2
Six *(Start, Mid1, Mid2, Mid3, End, Glob)*	53.7	35.9	8.1	35.1	18.1

four-phase design (start, middle, end, global) in our final model, which provides an effective trade-off while maintaining strong generalization, consistent with conclusions in the main submission on THUMOS14 dataset.

2.3. Analysis of Different LLM Backbone

We further extend the analysis of different LLM backbones beyond the THUMOS14 dataset under the 50% seen / 50%

Table 4. Analysis of Different LLM Backbone. The LLM backbone adopted in this paper is highlighted with a blue background.

Data Split	LLM Backbone	THUMOS14				ActivityNet v1.3			
		0.3	0.5	0.7	Avg	0.5	0.75	0.95	Avg
50% Seen 50% Unseen	Qwen3	64.9	49.1	24.0	46.2	53.3	35.2	7.5	34.3
	Deepseek v3	64.5	49.0	23.7	46.2	52.6	34.8	7.3	34.0
	GPT-4	65.1	49.4	24.1	46.6	52.9	35.3	7.6	34.5
	GPT-4o	65.4	49.7	24.3	46.9	53.1	35.3	7.7	34.6
75% Seen 25% Unseen	Qwen3	69.6	54.0	27.5	51.6	54.8	36.5	8.5	36.7
	Deepseek v3	69.8	53.7	27.5	51.4	55.1	36.7	8.3	37.0
	GPT-4	70.0	54.3	28.1	51.8	55.9	37.6	8.4	37.2
	GPT-4o	70.5	54.6	28.3	52.1	56.2	37.8	8.6	37.4

Please watch the video depicting a specific action, and evaluate the action's corresponding phase-wise descriptions according to the following criteria:

EVALUATION CRITERIA (Rate each criterion on a 1–10 scale, where10 is excellent):

1) **Linguistic Quality:** Whether the generated descriptions are grammatically correct and natural.

2) **Semantic Accuracy:** Whether the generated descriptions correctly reflect the true meaning and core semantics of each phase for the target action.

3) **Phase Clarity and Coherence:** Whether the phase decomposition is clear, and the descriptions across phases are temporally coherent and logically connected.

4) **Visual Alignment:** Whether the phase-wise descriptions align with the corresponding visual content.

5) **Transferability:** Whether the phase-wise descriptions of the current action exhibit similar or overlapping phase semantics to those of other actions.

Figure 1. Evaluation Template for GPT-4V and Human Assessments: Rating Across Five Dimensions—Linguistic Quality, Semantic Accuracy, Phase Clarity and Coherence, Visual Alignment, and Transferability.

Table 5. GPT-4V and Human Evaluation of Description Quality on THUMOS14. The best results are highlighted in **Red**.

Evaluation	Method	Rating [score/10.00]↑				
		Quality	Accuracy	Coherence	Alignment	Transferability
GPT-4V	Ground Truth	9.03	**8.35**	8.43	**8.15**	7.97
	Ours	**9.21**	8.22	**8.72**	7.86	**8.49**
Human	Ground Truth	**8.77** ±0.09	**8.10** ±0.15	8.23 ±0.10	**7.91** ±0.07	7.71 ±0.06
	Ours	8.46 ±0.06	7.95 ±0.08	**8.39** ±0.12	7.64 ±0.09	**8.28** ±0.10

unseen split in the main submission. Specifically, we additionally evaluate ActivityNet v1.3 dataset under both the 50% seen / 50% unseen and 75% seen / 25% unseen splits, as well as THUMOS14 under the 75% seen / 25% unseen split. The results on four widely used LLMs: Qwen3 [15], Deepseek v3 [10], GPT-4 [1], and GPT-4o [4] are shown in Table 4. we observe that overall performance across both datasets and evaluation splits remains relatively stable, regardless of the LLM backbone used. The relatively minor differences among the LLMs further suggest that our method is robust to the choice of LLM backbone, which is desirable for practical deployment.

2.4. Analysis of Description Quality

In this paper, we leverage the chain-of-thought (CoT) capability of LLMs to decompose action labels into coherent multi-phase descriptions. To evaluate the quality of these generated descriptions, we conduct both GPT-based and human-based subjective assessments. For GPT evaluation, we employ the multimodal LLM GPT-4V [17], and for human evaluation, we recruit ten volunteers. First, several

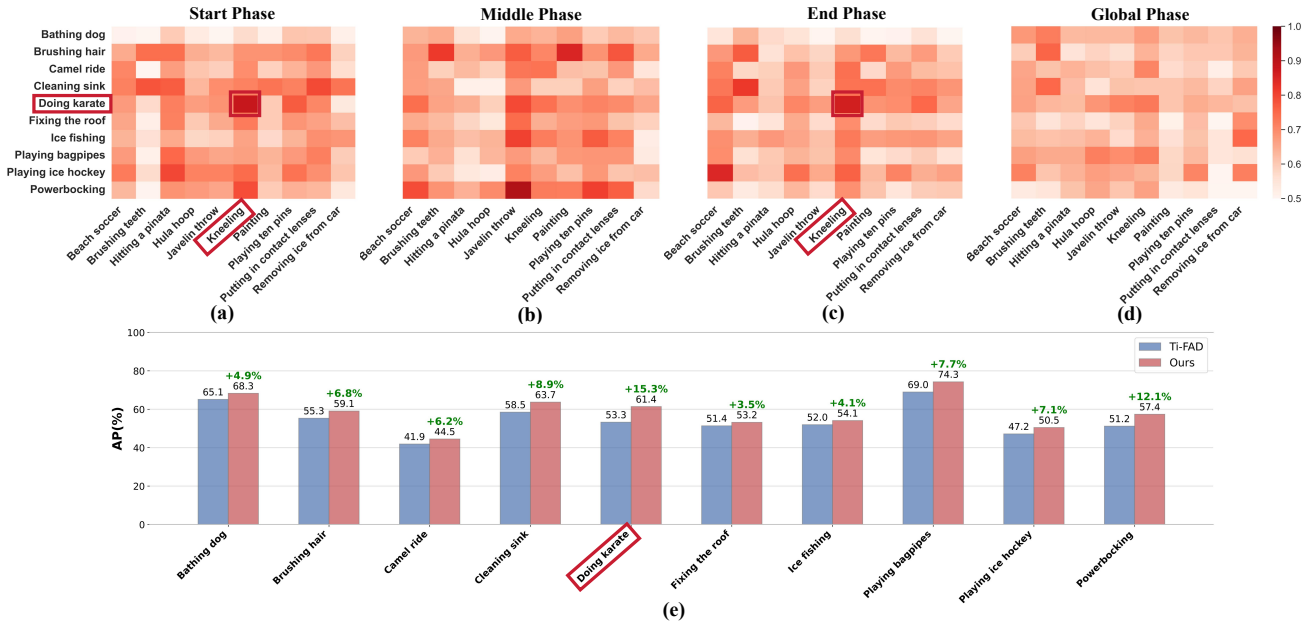

Figure 2. Phase-wise Semantic Similarity ((a)-(d)) and Per-unseen class AP (%) (e) at tIoU threshold 0.5 on ActivityNet v1.3 under the 50% seen / 50% unseen split. For (a)-(d), the vertical denotes unseen (testing) classes, the horizontal denotes seen (training) classes.

Table 6. Plug-and-Play Capability of the Proposed PDA on THUMOS14 under the 50% seen / 50% unseen split.

Backbone	Method	mAP@tIOU (%)			
		0.3	0.5	0.7	Avg
DyFADet (ECCV'24)	Baseline	17.5	12.2	5.7	11.9
	PDA	**66.1**	**50.2**	**25.0**	**47.5**
DiGIT (CVPR'25)	Baseline	19.1	13.5	6.1	13.0
	PDA	**67.3**	**50.7**	**25.2**	**48.0**

Table 7. Effectiveness of varying visual/textual backbones on THUMOS14.

Method	Feature		mAP@AVG	
	Visual	Text	50%-50%	75%-25%
Ti-FAD	CLIP-B	CLIP-B	27.3	29.7
	CLIP-L	CLIP-L	27.2	30.6
	I3D	CLIP-B	41.2	46.8
	I3D	CLIP-L	40.6	47.3
Ours	CLIP-B	CLIP-B	**30.1**	**33.5**
	CLIP-L	CLIP-L	**30.6**	**34.3**
	I3D	CLIP-B	**46.9**	**52.1**
	I3D	CLIP-L	**47.2**	**52.9**

domain experts are invited to decompose each action label into start, middle, end, and global descriptions, which serve as the ground truth. Subsequently, GPT-4V and human volunteers are employed to evaluate the generated descriptions across five dimensions: Linguistic Quality, Semantic Accuracy, Phase Clarity and Coherence, Visual Alignment, and Transferability. The detailed evaluation protocol is illustrated in Figure 1. For each action label, two corresponding videos are randomly selected for assessment. To alleviate potential scoring bias among human evaluators, we further compute the confidence levels of their ratings. The aggregated results are reported in Table 5. Experimental results show that the generated descriptions achieve comparable performance to human-decomposed ones in terms of linguistic quality, semantic accuracy, phase clarity, and visual alignment, indicating that the generated descriptions are linguistically natural and effectively capture the underlying action semantics. Notably, the generated descriptions score higher on transferability metric, indicating a stronger capacity to capture cross-action phase regularities learned from large-scale textual knowledge, which could further enhance zero-shot generalization in open-vocabulary temporal action detection.

Table 8. Comparison with Previous LLM-base Label Expansion Methods.

Task	Characteristics	Summary
Action Localization [2, 18]	Decompose actions into defining attributes and aggregate these attributes to align with frame-level embeddings, enabling more precise localization.	***Enrich textual semantics for more concise alignment.***
Action Recognition [3, 5]	Decompose actions into multi-dimensional descriptions and aggregate these descriptions to align with averaged visual embeddings for more precise recognition.	
Ours	Decompose actions into ***multi-phase descriptions*** and adaptively perform ***phase-wise alignment*** with visual features to learn transferable action patterns, enhancing zero-shot action detection performance.	***Learn transferable action knowledge for generalized zero-shot detection.***

Table 9. Comparison with LLM-base Label Expansion under the 50% seen / 50% unseen split.

Expansion	**Method**	**THUMOS14**				**ActivityNet v1.3**			
		0.3	0.5	0.7	Avg	0.5	0.75	0.95	Avg
Baseline	-	56.2	42.7	20.4	40.3	49.7	31.5	4.9	31.2
Global Label Expansion	(a) w/o Decompose	60.1	46.5	22.2	43.4	51.7	33.2	6.3	32.7
	(b) w/ Decompose	61.1	47.3	22.7	44.0	51.8	33.5	6.6	32.9
Single-phase Expansion	Start	57.4	43.8	21.1	41.2	50.1	32.0	5.2	31.6
	Middle	58.1	44.2	21.4	41.5	50.8	32.4	5.7	32.0
	End	57.8	44.6	21.5	41.7	50.6	32.5	5.3	31.8
	Global	59.3	45.5	21.9	42.5	51.0	32.8	5.9	32.3
Ours	-	**65.4**	**49.7**	**24.3**	**46.9**	**53.1**	**35.3**	**7.7**	**34.6**

3. More Experimental Results

3.1. Plug-and-Play Capability of the Proposed PDA

To assess the generalization and versatility of the proposed Phase-wise Decomposition and Alignment (PDA) framework, we conduct plug-and-play experiments on two recent closed-set TAD models, DyFADet [16] and DiGIT [7]. We integrate the PDA modules into these models and evaluate whether the introduced decomposition and alignment mechanisms improve their performance under the open-vocabulary setting. As reported in Table 6, our method consistently surpasses the baselines directly adapted to the OV-TAD protocol in [9, 14], achieving substantial gains in recognizing unseen action categories. These results demonstrate that PDA—through LLM-based multi-phase semantic decomposition followed by adaptive phase-wise alignment—effectively learns transferable action patterns and could serve as a versatile, plug-and-play component that enhances the generalization capability of existing closed-set TAD models in open-vocabulary scenarios.

3.2. Effects of Different Visual/Textual Backbones

In this section, we investigate the robustness of our approach under different visual and textual backbones. For the visual encoder, following [8], we consider CLIP ViT-B/16, ViT-L/14 and I3D. For the textual encoder, we consider CLIP ViT-B/16 and ViT-L/14. We compare against the global-alignment SOTA method Ti-FAD, which directly matches label-level semantics with global visual representations. As shown in Table 7, our method consistently surpasses Ti-FAD across all combinations of visual and textual encoders. This demonstrates the universality of our approach and its capacity to generalize across diverse backbone settings, thereby supporting improved open-vocabulary temporal action detection.

3.3. Difference Against Previous LLM-base Label Expansion Methods

As summarized in Table 8, some previous methods [2, 3, 6, 18] also use LLM to expand labels, but their goal is to generate more detailed semantic descriptions for visual matching. However, our approach leverages LLMs to extract transferable knowledge across semantically diverse labels. Combined with adaptive phase-wise alignment, this enables the discovery of phase-level transferable action patterns, thereby enhancing zero-shot detection. To clarify the source of our performance gain and distinguish our method from simple LLM-based label expansion, we design two comparison variants: 1) Global Label Expansion includes two sub-variants: (a) using GPT-4o to generate detailed ac-

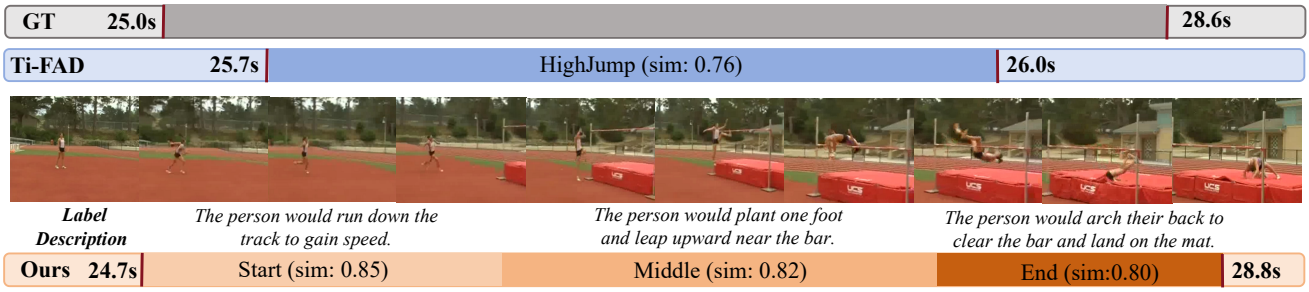

Figure 3. Visualization of the detection results "HighJump" on THUMOS14 under the 50% seen / 50% unseen split. "sim" represents the visual-textual similarity at each phase.

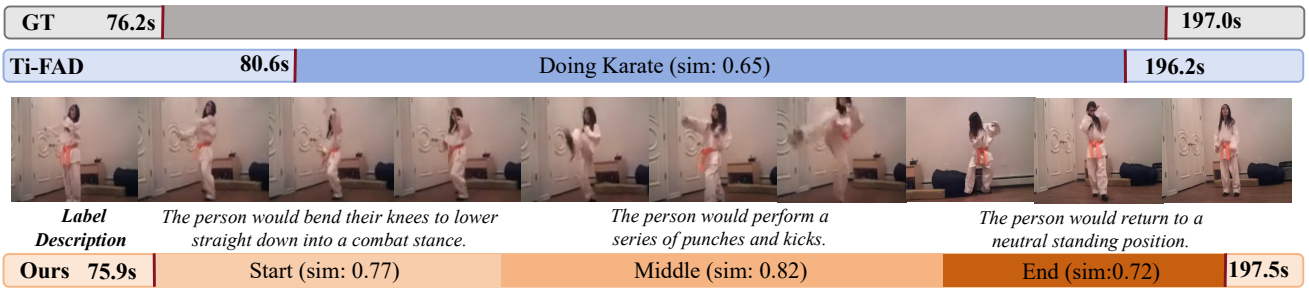

Figure 4. Visualization of the detection results "Doing Karate" on ActivityNet v1.3 under the 50% seen / 50% unseen split. "sim" represents the visual-textual similarity at each phase.

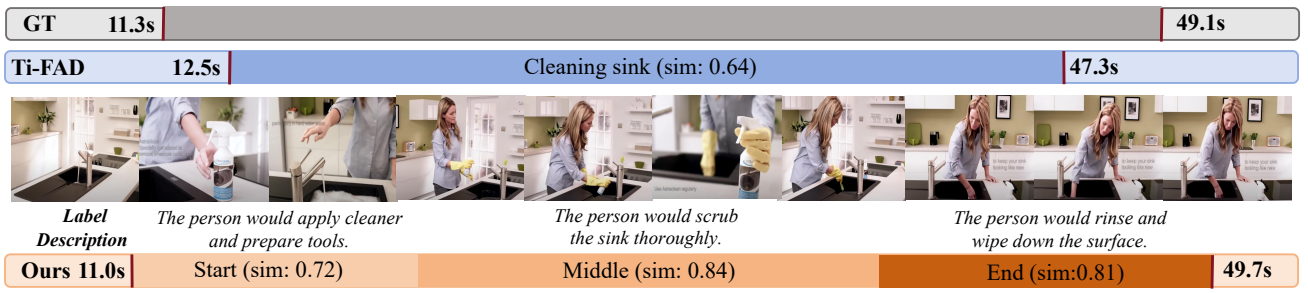

Figure 5. Visualization of the detection results "Cleaning sink" on ActivityNet v1.3 under the 50% seen / 50% unseen split. "sim" represents the visual-textual similarity at each phase.

tion descriptions without phase decomposition; (b) generating start, middle, end, and global phase descriptions with GPT-4o and concatenating them into a single expanded label. These enriched textual descriptions are aligned with video features using a global alignment strategy, similar to prior LLM-augmented methods. 2) Single-Phase Expansion leverages phase-specific semantic descriptions (Start, Middle, End, Global) to match video features individually. As shown in Table 9, global label expansion yields only marginal improvements, indicating that **richer semantics alone offer limited gains**. The single-phase variant performs even worse. In contrast, our full PDA framework achieves significant improvements, confirming that **the core advantage stems** not from single LLM-based label expansion but **from the combination with adaptive phase-wise visual–textual alignment**, which effectively transfers fine-grained visual priors from seen to unseen actions.

3.4. Phase-wise Semantic Similarity on Activitynet v1.3.

We further evaluate the effectiveness of PDA on ActivityNet v1.3 by analyzing phase-level semantic similarity and reporting per-class AP for 10 randomly selected seen and unseen classes. Figures 2 (a)-(d) present phase-wise semantic similarity matrices between seen and unseen classes, darker red regions indicate higher cosine similarity between cor-

responding phase descriptions. Notably, several unseen actions exhibit strong semantic alignment with seen actions at specific phases, despite differing at the global action level. A representative example is the seen-unseen pair Kneeling and Doing karate, which display high semantic similarity in both the start and end phases. Both actions begin with a similar preparatory motion—*"The person would bend their knees to lower straight down toward the ground"* (Kneeling) versus *"The person would bend their knees to lower straight down into a combat stance"* (Doing karate). Their ending phases also involve returning to an upright position. Such shared phase patterns provide transferable cues that our model effectively leverages during inference on unseen classes. In addition, Figure 2 (e) reports per-class Average Precision (AP) comparisons between our method and Ti-FAD. Consistent with results on THUMOS14, our approach achieves higher AP across all unseen categories, particularly those with strong phase-level semantic affinity to the seen set (e.g., Doing karate). These findings further validate that adaptive phase-aware decomposition promotes more effective knowledge transfer and enhances generalization to previously unseen actions.

3.5. More Qualitative Results

To further demonstrate the effectiveness of our proposed framework, we present additional qualitative results on one unseen action class (HighJump) from THUMOS14 and two unseen classes (Doing karate and Cleaning sink) from ActivityNet v1.3. As shown in Figure 3,4,5, our method consistently yields more accurate temporal boundaries and higher phase-level matching scores compared to the baseline Ti-FAD. Notably, the semantics of the corresponding phase-specific textual descriptions exhibit strong alignment with the predicted segments , indicating that our model not only improves classification accuracy but also enhances localization precision across datasets and action types. These consistent improvements further underscore the generalizability of our approach in recognizing unseen categories through fine-grained visual-semantic reasoning.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Josiah Aklilu, Xiaohan Wang, and Serena Yeung-Levy. Zero-shot action localization via the confidence of large vision-language models. *arXiv preprint arXiv:2410.14340*, 2024. 5

[3] Massimo Bosetti, Shibingfeng Zhang, Bendetta Liberatori, Giacomo Zara, Elisa Ricci, and Paolo Rota. Text-enhanced zero-shot action recognition: A training-free approach. In *International Conference on Pattern Recognition*, pages 327–342. Springer, 2024. 5

[4] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 3

[5] Chengyou Jia, Minnan Luo, Xiaojun Chang, Zhuohang Dang, Mingfei Han, Mengmeng Wang, Guang Dai, Sizhe Dang, and Jingdong Wang. Generating action-conditioned prompts for open-vocabulary video action recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4640–4649, 2024. 5

[6] Chen Ju, Zeqian Li, Peisen Zhao, Ya Zhang, Xiaopeng Zhang, Qi Tian, Yanfeng Wang, and Weidi Xie. Multi-modal prompting for low-shot temporal action localization. *arXiv preprint arXiv:2303.11732*, 2023. 5

[7] Ho-Joong Kim, Yearang Lee, Jung-Ho Hong, and Seong-Whan Lee. Digit: Multi-dilated gated encoder and central-adjacent region integrated decoder for temporal action detection transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24286–24296, 2025. 5

[8] Yearang Lee et al. Text-infused attention and foreground-aware modeling for zero-shot temporal action detection. *Advances in Neural Information Processing Systems*, 37:9864–9884, 2024. 1, 5

[9] Zhiheng Li, Yujie Zhong, Ran Song, Tianjiao Li, Lin Ma, and Wei Zhang. Detal: Open-vocabulary temporal action localization with decoupled networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):7728–7741, 2024. 5

[10] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 3

[11] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 1

[12] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Zero-shot temporal action detection via vision-language prompting. In *European conference on computer vision*, pages 681–697. Springer, 2022. 1

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[14] Song-miao Wang et al. Concept-guided open-vocabulary temporal action detection. *Journal of Computer Science and Technology*, 2025. 1, 5

[15] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3

[16] Le Yang, Ziwei Zheng, Yizeng Han, Hao Cheng, Shiji Song, Gao Huang, and Fan Li. Dyfadet: Dynamic feature aggrega-

tion for temporal action detection. In *European Conference on Computer Vision*, pages 305–322. Springer, 2024. 5

[17] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023. 3

[18] Minghang Zheng, Xinhao Cai, Qingchao Chen, Yuxin Peng, and Yang Liu. Training-free video temporal grounding using large-scale pre-trained models. In *European Conference on Computer Vision*, pages 20–37. Springer, 2024. 5